Marco Aiello
Ian E. Pratt-Hartmann
Johan F.A.K. van Benthem
Editors
Handbook of Spatial Logics

Spatial logics are used for describing and reasoning about any kind of geometrical entities and relations, broadly construed. In the past decade, such logics have attracted attention in such diverse fields as Artificial Intelligence, Computer Science, Mathematics, Philosophy and Physics. This handbook creates, for the first time, a systematic account of the resulting field. Following a general introduction, there are fourteen chapters by invited authors, with self-contained overviews, the principal insights and methods to date, and the central open problems.
The result is a comprehensive survey of this rapidly expanding subject.

Recently there has been a vigorous development in logics of space, stimulated by new applications in Computer Science and Artificial Intelligence, and by new attitudes to Logic which these disciplines have helped to foster. The Handbook will be a landmark publication, defining the field, and providing a superb resource for students and researchers, specialists and visitors alike.
SAMSON ABRAMSKY, Strachey Professor of Computing, Oxford University

Logic has been the language of choice for the study of Space ever since Euclid's Elements. This alliance has deeply influenced the way we do science. Contemporary Logic - with its new tools - continues this tradition, while adding dynamic and information-oriented themes to the mix. This Handbook covers the extensive new developments in spatial reasoning, and it may revolutionize the field.
SERGEI N. ARTEMOV, Professor of Computer Science and Mathematics, City University of New York

The geometry and dynamics of space have motivated much mathematics, but logic has lagged behind. This prodigious volume changes that picture significantly. It provides a wealth of material of wide interest, from physics to philosophy. Many chapters are substantial monographs in their own right. The book will stimulate further research deepening our understanding of the world around us.
ROBERT GOLDBLATT, Professor of Mathematics, Victoria University of Wellington

This book redraws boundaries in an imaginative way. Its common theme is how to represent spatial arguments within limited formalisms. Motivations for this range from image processing to foundations of physics, and from computational complexity to whole-part structures. Besides presenting many things that deserve to be better known, the book invites us to find the logical viewpoint that unifies them.
WILFRID HODGES, School of Mathematical Sciences, Queen Mary College, London

Logics of time have been developed over many decades. Logics of space appear to be much harder, and this book contains several new attempts which promise applications to Artificial Intelligence, especially to robotic manipulation of solid objects.
JOHN McCARTHY, Department of Computer Science, Stanford University

The precise relationship between spatial and logical structure has been an outstanding issue in Logic, Philosophy and Mathematics since the time of Immanuel Kant. This Handbook provides an up-to-date and accessible introduction to the current state of the art, in the light of recent work in Logic and Computer Science.
MICHAEL FRIEDMAN, Department of Philosophy, Stanford University

ISBN 978-1-4020-5586-7
9 781402 055867

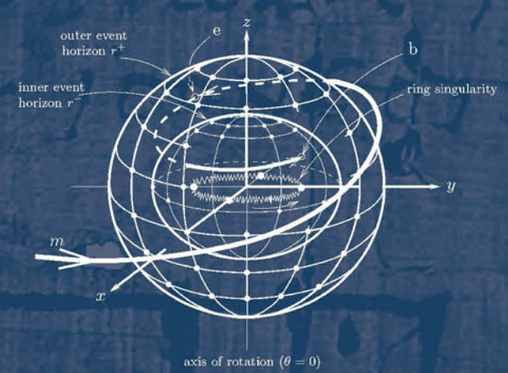
springer.com

Aiello
Pratt-Hartmann
van Benthem *Eds.*

Handbook of Spatial Logics

Marco Aiello
Ian E. Pratt-Hartmann
Johan F.A.K. van Benthem
Editors

Handbook of Spatial Logics



Springer

HANDBOOK OF SPATIAL LOGICS

HANDBOOK OF SPATIAL LOGICS

Edited by

MARCO AIELLO
University of Groningen

IAN PRATT-HARTMANN
University of Manchester

and

JOHAN VAN BENTHEM
University of Amsterdam

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-5586-7 (HB)
ISBN 978-1-4020-5587-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Cover Design by Lorenzo Rumori

Printed on acid-free paper

All Rights Reserved
© 2007 Springer
No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Preface

Space, with its manifold layers of structure, has been an inexhaustible source of intellectual fascination since Antiquity. The science that began with the empirical discoveries of the Egyptian 'rope-stretchers', and that has inspired many of the greatest developments in mathematics over the centuries, now comprises such topics as spatial databases, automated geometrical reasoning and digital image processing. In this long intellectual history, however, one relatively recent, yet crucial, event stands out: the rise of the logical stance in geometry. Fundamental to this development is the analysis of geometrical structures in relation to the formal languages used to describe them, and the recognition of the special mathematical challenges—and opportunities—which such an analysis presents. The interplay between logic and geometry is the subject of this book.

By a *spatial logic*, we mean any formal language for describing geometrical entities and configurations, where 'geometrical' is understood in a broad sense. Unlike their well-studied temporal counterparts, spatial logics have been curiously neglected in the literature on mathematical logic, despite some early pioneering work by Tarski and others on the foundations of geometry and topology in the middle years of the previous century. Only in the last decade have spatial logics attracted renewed attention from logicians, partly as a response to work in such diverse fields as artificial intelligence, database theory, physics and philosophy.

Today, there is a fast-growing body of literature on spatial logics, with motivations ranging from computational issues to the foundations of knowledge and information. But, for the newcomer to the field, this diversity of influences and approaches constitutes something of a mixed blessing: the field may be in a state of rapid development; but there is as yet no common research agenda, and no common vocabulary to allow ideas to be shared across disciplines. The aim of this book is to provide a resource which presents a view of the best current work in different communities worldwide, and which makes a first attempt at systematic linkage. We hope this will stimulate the development of spatial logic itself, but beyond this narrower purpose, we also hope to have provided a text that should be of value to non-logicians with an interest in formal theories of space.

The book consists of a general introduction followed by fourteen invited contributions on various topics in spatial logic, with authors representing the major active centres in the field. Each of these chapters provides a self-contained overview of its topic, describing the principal results obtained to date, explaining the methods used to obtain them, and listing the most important open problems.

Every contributed chapter has one or more 'second readers'—experts in the field, who worked with the authors and editors to help ensure a comprehensive (and comprehensible) account of the topic in question.

The book is intended as a technical resource for academic researchers and graduate students. Familiarity with basic undergraduate-level logic, topology and geometry is generally assumed. Very roughly, the criterion of accessibility we have worked to is that a good graduate student interested in the area should, by working through any of these chapters, be able to acquire a firm understanding of the current state-of-the-art in that topic within the space of a few weeks. Jointly, these chapters provide—to the extent that this is ever possible in a rapidly evolving discipline—a comprehensive survey of the field of spatial logic as it stands today.

Marco Aiello

Ian Pratt-Hartmann

Johan van Benthem

Contents

Preface v

Contributing Authors xi

Second Readers xxi

1
What is Spatial Logic? 1
Marco Aiello, Ian Pratt-Hartmann, Johan van Benthem

2
First-Order Mereotopology 13
Ian Pratt-Hartmann
1. Introduction 13
2. Mereotopologies 14
3. Defining topological relations 26
4. Expressiveness of first-order languages in plane mereotopologies 38
5. Axiomatization 58
6. Spatial mereotopology 69
7. Model theory 82
8. Philosophical considerations 91

3
Axioms, Algebras and Topology 99
Brandon Bennett, Ivo Düntsch
1. Introduction 99
2. Preliminary definitions and notation 104
3. Contact relations 119
4. Boolean contact algebras 122
5. Other theories of topological relations 134
6. Reasoning about topological relations 137
7. Conclusion 149

4
Qualitative Spatial Reasoning Using Constraint Calculi 161
Jochen Renz, Bernhard Nebel
1. Introduction 161
2. Constraint-based methods for qualitative spatial representation and reasoning 163
3. Spatial constraint calculi 169

4. Computational complexity 177
5. Identifying tractable subsets of spatial CSPs 184
6. Practical efficiency of reasoning methods 190
7. Combination of spatial calculi 197
8. Conclusions 207

5
Modal Logics of Space 217
Johan van Benthem, Guram Bezhanishvili
1. Modal logics and spatial structures 217
2. Modal logic and topology: basic results 231
3. Modal logic and topology: further directions 256
4. Modal logic and geometry 276
5. Modal logic and linear algebra 285
6. Conclusions 291

6
Topology and Epistemic Logic 299
Rohit Parikh, Lawrence S. Moss, Chris Steinsvold
1. Introduction 299
2. Perspectives 300
3. The original topological interpretation of modal logic: Tarski and McKinsey's Theorem 301
4. Topologic 308
5. A logical system: the subset space axioms 312
6. Further examples 316
7. Completeness of the subset space axioms 319
8. Decidability of the subset space logic 322
9. Heinemann's extensions to topologic 325
10. Common knowledge in topological settings 327
11. The topology of belief 329
12. Other work connected to this chapter 339

7
Logical Theories for Fragments of Elementary Geometry 343
Philippe Balbiani, Valentin Goranko, Ruaan Kellerman, Dimiter Vakarelov
1. Introduction and historical overview 343
2. Preliminaries 348
3. Structures and theories of parallelism 352
4. Structures and theories of orthogonality 355
5. Two-sorted point-line incidence spaces 358
6. Coordinatization 367
7. On the first-order theories of affine and projective spaces 380
8. Betweenness structures and ordered affine planes 386
9. Rich languages and structures for elementary geometry 394
10. Modal logic and spatial logic 400
11. Point-based spatial logics 404

12. Line-based spatial logics 406
13. Tip spatial logics 411
14. Point-line spatial logics 416

8
Locales and Toposes as Spaces 429
Steven Vickers
1. Introduction 429
2. Opens as propositions 431
3. Predicate geometric logic 445
4. Categorical logic 457
5. Sheaves as predicates 474
6. Summary of toposes 488
7. Other directions 489
8. Conclusions 492

9
Spatial Logic + Temporal Logic = ? 497
Roman Kontchakov, Agi Kurucz, Frank Wolter, Michael Zakharyaschev
1. Introduction 497
2. Static and changing spatial models 501
3. Spatial logics 506
4. Temporal logics 527
5. Combination principles 531
6. Combining topo-logics with temporal logics 533
7. Combining distance logics with temporal logics 543
8. Logics for dynamical systems 546
9. Related 'temporalised' formalisms 557

10
Dynamic Topological Logic 565
Philip Kremer, Grigori Mints
1. Introduction 565
2. Basic definitions 569
3. Recurrence and the DTL of measure-preserving continuous functions on the closed unit interval 573
4. Purely topological and purely temporal fragments of DTLs 576
5. S4 is topologically complete for (0, 1) 579
6. The logic of homeomorphisms 586
7. The logic of continuous functions 592
8. Conclusion 604

11
Logic of Space-Time and Relativity Theory 607
Hajnal Andréka, Judit X. Madarász, István Németi
1. Introduction 607
2. Special relativity 608
3. General relativistic space-time 660

4. Black holes, wormholes, timewarp. Distinguished general relativistic space-times 683
5. Connections with the literature 705

12
Discrete Spatial Models 713
Michael B. Smyth, Julian Webster
1. Introduction 713
2. Preliminaries; correspondence principle 717
3. Čech closure spaces 723
4. Closure systems 725
5. Extended examples 730
6. (Boundary and) dimension 743
7. Discrete region geometry 749
8. Matroids 761
9. Spherical oriented matroids 768
10. Flat oriented matroids 783
11. Algebraic spatial models 787

13
Real Algebraic Geometry and Constraint Databases 799
Floris Geerts , Bart Kuijpers
1. From the relational database model to the constraint database model 799
2. Constraint data models and query languages 805
3. Introduction to real algebraic geometry 812
4. Query evaluation through quantifier elimination 822
5. Expressiveness results 829
6. Extensions of logical query languages 841

14
Mathematical Morphology 857
Isabelle Bloch, Henk Heijmans, Christian Ronse
1. Introduction 857
2. Algebra 876
3. Related approaches 897
4. Logics 918
5. Conclusion 936

15
Spatial Reasoning and Ontology: Parts, Wholes, and Locations 945
Achille C. Varzi
1. Philosophical issues in mereology 947
2. Philosophical issues in topology 975
3. Location theories 1012

Index 1039

Contributing Authors

Marco Aiello is Professor of Distributed Information Systems at the University of Groningen. He holds an MSc. in Engineering and Computer Science from the University of Rome "La Sapienza" (1997) and a PhD. in Computer Science and Logic from the University of Amsterdam (2002). From 2002 to 2005 he was Assistant Professor at the University of Trento, while in 2006 he was a Lise Meitner fellow at the Vienna University of Technology. Dr. Aiello's research interests revolve around the notion of space, including modal logics of space, document understanding via spatial interpretation, spatially distributed systems and service-oriented computing.

Hajnal Andréka is Head of the Department of Algebraic Logic at the Rényi Mathematical Research Institute, Budapest. She holds a Mathematics Diploma from the Eötvös University, Budapest, as well as the degree of Doctor of Mathematics from the Hungarian Academy of Sciences. Since 1977, she has been working in the Mathematical Research Institute of the Hungarian Academy of Sciences. She has published approximately 100 research papers and several books. Her main research interests centre around the connections between algebra, logic and geometry, in particular: relativity theory, its logical analysis and foundations, spacetime theories, black holes and cosmology. Dr. Andréka is a member of various editorial boards of international journals and other international academic bodies. She has received numerous awards in Hungary in recognition of her work in mathematics.

Philippe Balbiani is Researcher in the Institut de Recherche en Informatique de Toulouse, where he leads the Logic, Interaction, Language and Computation group. He earned his PhD. in Computer Science from the Université Paul-Sabatier, Toulouse III, in 1991. He has worked principally on logic programming, qualitative reasoning and applied non-classical logics. Dr. Balbiani's current focus is on mathematical methods, models and architectures for computer security.

Brandon Bennett is Lecturer in the School of Computing at the University of Leeds. He has a BSc. in Physics and Computer Science, an MA. in Philosophy and a PhD. in Computer Science, all from the University of Leeds. His publications cover various subfields of knowledge representation, including spatial reasoning, foundational ontology, representing geographic information and logical modelling of vagueness. His current research focus is on the representation and manipulation of vague geographic objects within GIS applications.

Guram Bezhanishvili is Associate Professor in the Department of Mathematical Sciences at New Mexico State University. He obtained his PhD. in 1998 from the Tokyo Institute of Technology. Before moving to New Mexico State University, he held positions as Assistant Professor at the Tbilisi State University, and Postdoctoral Fellow in the Institute of Logic, Language and Information at the University of Amsterdam. His main interests lie in the use of algebraic and topological methods in non-classical logics.

Isabelle Bloch is Professor in the Signal and Image Processing Department (*Département TSI*) at the École Nationale Supérieure des Télécommunications. She obtained her PhD. in 1990 and her Habilitation diploma in 1995. Her research interests include 3D image and object processing, 3D and fuzzy mathematical morphology, decision theory, information fusion, fuzzy set theory, belief function theory, structural pattern recognition, spatial reasoning and medical imaging.

Ivo Düntsch is Professor of Computer Science at Brock University. He obtained his PhD. from the Free University of Berlin in 1981. He worked as a lecturer at Bayero University in Kano, Nigeria, and was founding chair of the Mathematics Department at the University of Brunei Darussalam. From 1991 until 1994, he was Deputy Director of the Computer Centre at the University of Osnabrück, and from 1994 until 2002, held a Chair in Computer Science at the University of Ulster. His research interests lie in the area of the logical and algebraic foundations of non–invasive data analysis, in particular, rough set theory and qualitative spatial reasoning.

Floris Geerts is Post-Doctoral Researcher of the Research Foundation of Flanders at Hasselt University and the Transnational University of Limburg, and a Research Associate in the Database Group at the University of Edinburgh. He has a Master's degree in Mathematics from the University of Ghent and a Doctoral degree in Computer Science from Hasselt University. He spent two years as a Post-Doctoral Researcher at the University of Helsinki. His interests

include the study of constraint databases, query languages in the context of XML and, more recently, data models and query languages for annotated scientific data. He has a keen interest in connections between geometry and logic.

Valentin Goranko is Associate Professor in the School of Mathematics at the University of the Witwatersrand. He obtained his Master's and PhD. degrees in Mathematics (mathematical logic) from the University of Sofia, and then held academic positions at the University of Sofia, the Rand Afrikaans University and the University of the Witwatersrand. He has over 40 research publications, mainly in theory and applications of modal and temporal logics to computer science and artificial intelligence. His current interests also include logical theories of geometric structures, logics of multi-agent systems and logic computation in finitely presentable infinite structures.

Henk Heijmans is the director of the "Signals and Images" research theme at the CWI in Amsterdam. He received his Master's degree in Mathematics from the Technical University of Eindhoven and his PhD. from the University of Amsterdam in 1985. His research interests are focused towards mathematical techniques for image and signal processing, with an emphasis on mathematical morphology and wavelet analysis.

Ruaan Kellerman is a PhD. student in the School of Mathematics at the University of the Witwatersrand. He holds an MSc. from the Department of Mathematics at the University of Johannesburg, the thesis topic of which was the logical theories of geometric orthogonality structures. His research currently involves the logical theories of trees.

Roman Kontchakov is Postdoctoral Research Fellow in the School of Computer Science and Information Systems at Birkbeck College, London. He received his MSc. in Applied Mathematics (with Honours) from Moscow State University in 1999 and his PhD. from King's College, London in 2004. His research interests include first-order, modal and temporal logics, description, metric and spatial logics, combinations of logics, decidability and computational complexity of logics, web services and the Semantic Web.

Philip Kremer is Associate Professor in the Department of Philosophy at the University of Toronto. He has a BSc. in Mathematics from the University of Toronto (1985) and a PhD. in Philosophy from the University of Pittsburgh (1994). He was an Assistant Professor of Philosophy at Stanford University (1994-1996), an Assistant Professor at Yale University (1996-1999) and an

Associate Professor of Philosophy at McMaster University (1999-2003). Since 2003 he has held an undergraduate appointment as an Associate Professor of Philosophy in the Department of Humanities at the University of Toronto at Scarborough, together with a graduate appointment in the Department of Philosophy at the University of Toronto. He has published on dynamic topological logic, on truth and paradox, on propositional quantification, and on relevance logic.

Bart Kuijpers is Associate Professor in the Theoretical Computer Science group at Hasselt University and the Transnational University of Limburg. He has a Master's degree in Mathematics from the University of Leuven and a Doctoral degree in Computer Science from the University of Antwerp. He was a Researcher at the Universities of Leuven and Antwerp and was a Post-Doctoral Researcher of the Research Foundation of Flanders before becoming Professor of Theoretical Computer Science at Hasselt University. Since 2003, he has also been Visiting Researcher at the University of Buenos Aires. His main research is in query evaluation and the expressive power of database query languages for (possibly infinite) database systems that are described by constraints. His interests also include data models and query languages for spatio-temporal data.

Agi Kurucz is Senior Lecturer in the Department of Computer Science at King's College, London. She obtained her Diploma in Mathematics from the Eötvös University, Budapest in 1985, and her PhD. in Mathematics from the Hungarian Academy of Sciences in 1998. She lectured in the Department of Symbolic Logic at the Eötvös University, Budapest from 1997 to 1998. From 1998 to 2000, she worked as Research Associate in the Department of Computing at Imperial College, London. She joined King's College in 2001. Her main research interests are classical predicate logic and modal and algebraic logics.

Judit Madarász is Junior Research Fellow at the Rényi Mathematical Research Institute, Budapest. She received her Master's degree in Mathematics in 1995 and her PhD. in Mathematics in 2003, both from the Eötvös University, Budapest. She is a co-author of the internet book *Logical Structure of Relativity Theories*, and has published in numerous mathematical journals. Dr. Madarász' main research interests include relativity theory (both special and general), spacetime, logical foundations of relativity theories and geometry, black holes, cosmology and algebraic logic.

Grigori Mints is Professor of Philosophy and (by courtesy) of Mathematics and Computer Science at Stanford University. He received his MSc. in

Mathematics in 1961, his PhD. in 1965 and his ScD. in 1990 (also in Mathematics) from St. Petersburg University. He has held appointments at the Steklov Institute of Mathematics, St. Petersburg, St. Petersburg University, the Institute of Cybernetics, Tallinn and visiting appointments in Amsterdam, Stockholm, Berkeley and Munich. He is the author and editor of 9 books, more than 200 papers and more than 2500 published reviews. His main research interests are logic and the foundations of mathematics. Prof. Mints is an editor of reviews for the *Bulletin of Symbolic Logic* and a member of the editorial boards of the *Journal of Philosophical Logic*, the *Journal of Logic and Computation*, and the *Logic Journal of IGPL.*

Lawrence Moss is the Director of the Indiana University Program in Pure and Applied Logic. He is also Professor of Mathematics and Adjunct Professor of Computer Science, Informatics, Linguistics and Philosophy. His PhD. was in Mathematics from UCLA in 1984. He has also held positions at Stanford University's Center for the Study of Language and Information, the University of Michigan and the IBM T. J. Watson Research Center. His research areas are mainly in applied logic, and include coalgebra, epistemic logic and interactions of logic and linguistics. Prof. Moss chairs the Steering Committee of the North American Summer School in Logic, Language and Information. He also serves on the editorial boards of a number of journals in his fields and on many conference programme committees.

Bernhard Nebel is Professor in the Department of Computer Science at the Albert-Ludwig University of Freiburg. He received the degree of Dipl.-Inform. from the University of Hamburg, and his PhD. (Dr. rer. nat.) from the University of the Saarland. Between 1982 and 1993 he worked on different AI projects at the University of Hamburg, the Technical University of Berlin, the Information Sciences Institute (University of Southern California), IBM Germany, and the German Research Center for AI (DFKI). From 1993 to 1996 he held an Associate Professor position at the University of Ulm. Since 1996 he has been head of the research group on Foundations of AI at Freiburg. His current research interests are action planning, robotics and temporal and spatial reasoning. Prof. Nebel is an ECCAI fellow, and has chaired various international conferences in artificial intelligence.

István Németi is Senior Scientific Advisor at the Rényi Mathematical Research Institute, Budapest. He received his Master's in Electrodynamics and his PhD. in Mathematics (1978), both from the Eötvös University, Budapest, and his Dr. Sci. from the Hungarian Academy of Sciences in 1987. Since 1974, he has been a researcher at the Mathematical Institute of the Hungarian

Academy of Sciences, and has taught at the Eötvös University. He is the co-author of several scientific books and approximately 130 research papers in leading scientific journals. Prof. Németi's research interests include logic and the boundaries between logic, geometry and algebra, as well as relativity theory and cosmology.

Rohit Parikh is Distinguished Professor of Computer Science at Brooklyn College, CUNY, also attached to the doctoral programmes of Computer Science, Mathematics and Philosophy at the City University Graduate Center. He received his doctorate from Harvard in 1962 with a dissertation on Transfinite Progressions. He is a three-times winner in the William Lowell Putnam mathematical competition. He has taught at Stanford University, Panjab University, Bristol University, SUNY, Buffalo and NYU. Before coming to City University he was Professor of Mathematics at Boston University for 15 years (the first five as Associate Professor). He has also worked at Bell Labs, IBM, the Tata Institute-Mumbai, Caltech and the ETH-Zurich. Prof. Parikh has published more than a hundred papers, mostly, but not entirely, in logic and its applications. His fields of interest include formal languages, theory of proofs, non-standard analysis, dynamic logic, logic of knowledge, game theory, philosophical logic and social software. He has been editor of the *International Journal for the Foundations of Computer Science* and the *Journal of Philosophical Logic.*

Ian Pratt-Hartmann is Senior Lecturer in the School of Computer Science at the University of Manchester. He read Mathematics and Philosophy at Brasenose College, Oxford, and Philosophy at Princeton University, receiving his PhD. there in 1987. Dr. Pratt-Hartmann has published widely in logic, cognitive science and artificial intelligence. His current research interests include (besides spatial logic) the complexity of decidable fragments of logic and the relationship between natural language and logic. He is a member of the editorial board of the *Journal of Logic, Language and Information.*

Jochen Renz is Fellow of the Research School of Information Sciences and Engineering at the Australian National University. He received his PhD. from the Albert-Ludwig University of Freiburg, and was a Postdoctoral Fellow at the Wallenberg Laboratory for Information Technology and Autonomous Systems at the University of Linköping. After a two-year Marie Curie Postdoctoral Fellowship at the Vienna University of Technology, he moved to National ICT Australia in Sydney. Dr. Renz' main research interests are in qualitative spatial and temporal representation and reasoning, in particular in the computational properties of reasoning and in efficient reasoning algorithms.

Christian Ronse is Professor of Computer Science at the Université Louis Pasteur, Strasbourg I, and member of the LSIIT (UMR 7005 CNRS-ULP) laboratory. He studied Pure Mathematics at the Free University of Brussels (Licence, 1976) and the University of Oxford (MSc., 1977; PhD., 1979), specialising in group theory. Between 1979 and 1991, he was a member of scientific staff at the Philips Research Laboratory, Brussels, where he conducted research on combinatorics of switching circuits, feedback shift registers, discrete geometry, image processing and mathematical morphology. During the academic year 1991–1992, he worked at the Université Bordeaux I, where he obtained his Habilitation diploma. Since October 1992, he has been Professor at Strasbourg, where he has contributed to the development of a research group on image analysis, and the teaching of image processing to students at various levels. His scientific interests include lattice theory, mathematical morphology, image segmentation and medical imaging. Prof. Ronse is a member of the editorial board of the *Journal of Mathematical Imaging and Vision*.

Mike Smyth recently retired from a Readership in Theoretical Computer Science at Imperial College, London, and currently holds an honorary position at the University of Birmingham. He studied Mathematics, Philosophy and Computer Science at several UK universities, and was awarded a D. Phil. in Mathematics at the University of Oxford in 1980. Dr. Smyth is probably best known for contributions to domain theory and asymmetric topology. In recent years his interests have shifted towards digital topology and, especially, to the task of developing geometry in such a way as to allow that space (or spacetime) is discrete.

Chris Steinsvold is Adjunct Professor at Brooklyn College, CUNY and the Borough of Manhattan Community College, CUNY. He received his BA in Philosophy from Brooklyn College, CUNY, and is currently working on his dissertation at the CUNY Graduate Center.

Dimiter Vakarelov is Professor of Logic in the Faculty of Mathematics and Informatics at the University of Sofia. He obtained his Master's degree in Mathematics from the University of Sofia, his PhD. degree in Mathematical Logic from the University of Warsaw and the degree of Doctor of Mathematical Sciences from the University of Sofia. He has more than 70 research publications. His scientific interests are in the field of non-classical logic (mainly modal logic) with applications in computer science and AI.

Johan van Benthem is University Professor of Logic at the University of Amsterdam, Professor of Philosophy at Stanford University, and a member of the informatics section of the Academia Europaea. He studied physics, philosophy, and mathematics at the University of Amsterdam (PhD. 1977), and was the founding director of the Institute for Logic, Language and Computation. He is co-editor of the *Handbook of Logic and Language* (1997) and the *Handbook of Modal Logic* (2006). His interests include modal logic, logics of time and space, dynamic logics of computation and information, and logic and game theory.

Achille Varzi is Professor of Philosophy at Columbia University. A graduate of the University of Trento, he received his PhD. in Philosophy from the University of Toronto. He is author or co-author of over 100 articles and several books in logic, metaphysics, and the philosophy of language. Prof. Varzi is currently an editor of the *Journal of Philosophy*, a subject editor of the *Stanford Encyclopedia of Philosophy*, and an associate editor of *Studia Logica*, *The Monist*, *Dialectica* and *Applied Ontology*.

Steve Vickers is Senior Lecturer in the School of Computer Science at the University of Birmingham. His career has combined pure mathematics and computing. After studying Mathematics at Cambridge and Leeds (PhD. in ring theory 1979), he worked on the Sinclair ZX81 and ZX Spectrum microcomputers, subsequently cofounded the company that developed the Jupiter Ace microcomputer. In 1985 he returned to mathematics at the Department of Computing, Imperial College, staying until 1999. He then moved to the Department Pure Mathematics at the Open University, and in 2003 he took up his present post at Birmingham. His main interests lie in the application of logic and category theory to topology.

Julian Webster studied Philosophy at Newcastle University, and obtained his Ph. D. under Mike Smyth at Imperial College, London, on the subject of the digital approximation of topologies and measures. He was subsequently a postdoctoral Research Associate at Imperial College.

Frank Wolter is Professor for Logic and Computation in the Department of Computer Science at the University of Liverpool. He received his PhD. in Mathematics from the Free University of Berlin in 1993, and his Habilitation in Computer Science at Leipzig University in 2000. His main interests are in knowledge representation and reasoning, in particular, description and spatial logics, and in modal and temporal logic. He is co-author and co-editor of

various monographs and collections. Prof. Wolter is a Fellow of the British Computer Society.

Michael Zakharyaschev is Professor of Computer Science in the School of Computer Science and Information Systems at Birkbeck College, London. He obtained his Diploma in Mathematics at Moscow State University in 1978, and his PhD. and Habilitation in Mathematics at Novosibirsk State University in 1985 and 1998. He was a Research Associate at the Keldysh Institute for Applied Mathematics, Russian Academy of Sciences, from 1978 to 1999, and lectured in the Faculty of Computational Mathematics and Cybernetics at Moscow State University from 1986 until 1998. He joined Birkbeck College in 2005. His main research interests are modal logic and its applications, description logic, spatial logic, classical predicate and intuitionistic logics and knowledge representation and reasoning. He is a co-author of two major research monographs on modal logic and the mathematical foundations and applications of many-dimensional modal logics.

Second Readers

Marco Aiello, *University of Groningen.*

Philippe Balbiani, *Institut de Recherche en Informatique de Toulouse.*

Guram Bezhanishvili, *New Mexico State University.*

Ivo Düntsch, *Brock University.*

Jen Davoren, *University of Melbourne.*

Antony Galton, *University of Exeter.*

István Németi, *Rényi Mathematical Research Institute, Budapest.*

Ian Pratt-Hartmann, *University of Manchester.*

Peter Revesz, *University of Nebraska–Lincoln.*

Darko Sarenac, *Stanford University.*

Valentin Shehtman, *Moscow State University and King's College, London.*

John Stell, *University of Leeds.*

Johan van Benthem, *University of Amsterdam and Stanford University.*

Yde Venema, *University of Amsterdam.*

Chapter 1

WHAT IS SPATIAL LOGIC?

Marco Aiello
University of Groningen

Ian Pratt-Hartmann
University of Manchester

Johan van Benthem
University of Amsterdam & Stanford University

By a *spatial logic*, we understand any formal language interpreted over a class of structures featuring geometrical entities and relations, broadly construed. The formal language in question may employ any logical syntax: that of first-order logic, or some fragment of first-order logic, or perhaps higher-order logic. The structures over which it is interpreted may inhabit any class of geometrical 'spaces': topological spaces, affine spaces, metric spaces, or perhaps a single space such a the projective plane or Euclidean 3-space. And the non-logical primitives of the language may be interpreted as any geometrical properties or relations defined over the relevant domains: topological connectedness of regions, parallelism of lines, or perhaps equidistance of two points from a third. What all these logics have in common is that the operative notion of validity depends on the underlying geometry of the structures over which their distinctively spatial primitives are interpreted. Spatial logic, then, is simply the study of the family of spatial logics, so conceived.

An analogy will help elucidate this rather austere-looking definition. From our stance, spatial logic parallels the more established area of temporal logic. A temporal logic is a formal language interpreted over some class of structures based on frameworks of temporal relations, broadly construed. The language in question, though usually some modal fragment of first- or higher-order logic, may in principle employ any logical syntax; the objects over which that syntax is interpreted may include points, paths, or extended intervals over any variety

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 1–11.
© 2007 *Springer.*

of partial orders; and the assumed partial order ultimately provides the interpretation for the distinctively temporal primitives of the formal language. What all temporal logics have in common, whether point- or interval-based, is that their operative notion of validity depends on the assumed properties of the underlying temporal flow. And what gives them their enduring appeal is the way in which the formal languages they employ balance expressive power against computational complexity. In this respect, temporal logic is the computationally motivated study of time.

Let us set the scene for the treatment of spatial logic in this book by examining some of the historical trends that have given rise to it. Classical geometry, the cultural model of deductive proof par excellence since Euclid's *Elements*, was finally analyzed in full mathematical precision in Hilbert's *Grundlagen der Geometrie* (Hilbert, 1909; see also Hilbert, 1950), when all its axioms, and possible variations on them, had become clear. Yet, despite its starkly abstract view of points, lines and planes, the *Grundlagen* is still couched not in a formal language, but rather in (lightly mathematicized) idiomatic German. Hilbert's Axiom of Parallels provides a good example:

> Let a be a line, and A a point not on a. Then, in the plane determined by a and A, there is at most one line which passes through A and does not meet a. (tr. from Hilbert, 1909, p. 20)

No attempt is made to tease out the implicit logical syntax of this language, or to analyze the underlying inference engine much beyond what Euclid had already done in his Common Notions. This is perhaps clearest in the case of Hilbert's final Axiom of Completeness:

> The elements (points, lines, planes) of the geometry form a system of objects which is not capable of any extension, subject to maintenance of all the preceding axioms. That is to say: it is not possible to add to the system of points, lines and planes another system of objects in such a way that, in the combined system, all [previous] axioms are satisfied. (*Ibid.*, p. 22.)

It was not until after the development of the apparatus of formal logic and model-theoretic semantics in the first half of the Twentieth Century that logicians were able to probe the precise inferential and expressive resources of geometry, in a second round of formalization culminating in Tarski's *Elementary Geometry* (Tarski, 1959).

Tarski's decisive contribution in his 1959 paper was not simply to force Hilbert's axioms into the regimented syntax of some formal language, but rather, to investigate what happens when that syntax is restricted. Specifically, Tarski employs a first-order logic, with variables ranging over points in the Euclidean plane, and with non-logical predicates standing for two primitive spatial relations: a ternary relation of 'betweenness' and a quaternary relation of 'equidistance'. The resulting language is sufficiently expressive to formulate much of Euclidean geometry—for example, Pythagoras' theorem, or the

existence of the nine-point Feuerbach circle. The computational reward for this loss of expressive power is considerable. Tarski showed that the theory of elementary geometry is *decidable*: there is a mechanical procedure to determine, of any given sentence in the relevant language, whether that sentence is true under the advertised interpretation. By contrast, the second-order theory needed to express all of Hilbert's axioms is undecidable.

Tarski's discovery illustrates the most distinctive feature of logic in the wake of the model-theoretic revolution of the previous century: its fundamentally linguistic orientation. The model-theoretic approach to logic takes as its central concern the often intricate relationship between mathematical structures and languages which describe them. On this view, spatial logic, as defined above, becomes the study of the relationship between geometrical structures and the spatial languages which describe them. It is this preoccupation with language which divides spatial logic from geometry as traditionally conceived. More recently, of course, the enterprise of automating logical deduction using electronic computers has necessitated new levels of precision and sophistication in reasoning about the properties of formal languages and their relationship to their subject matter. In this setting, the issue of balancing the expressive power of a language against the computational complexity of performing deductions within it occupies centre-stage.

We can broaden our perspective by considering two further examples of spatial logics in addition to Tarski's Elementary Geometry. To motivate our second example, recall that, in Elementary Geometry, all variables are taken to range over *points* in the Euclidean plane. This allows for quantification over geometrical figures defined by a fixed number of points, such as line segments, triangles, circles, and so on, but not over spatial constellations defined by point-sets of arbitrary finite size, such as polygons, let alone those defined by infinite sets of points, such as, for example, arbitrary connected regions. The question therefore arises as to what happens when these restrictions are lifted. In fact, Tarski himself had already investigated such a language in his *Geometry of Solids* (Tarski, 1956). This system employs the syntax of *second*-order logic, with the object variables ranging not over points, but instead over certain 'regions' in three-dimensional Euclidean space (hence, the set-variables range over sets of regions). The regions in question—Tarski called them *solids*—are the regular closed subsets of $\mathbb{R}^3$, namely, those subsets of $\mathbb{R}^3$ equal to the closure of their interior. Tarski's language features two non-logical predicates: one standing for the binary relation of *parthood*, the other for the unary property of *being spherical*. Again, Tarski establishes a remarkable fact about the relationship between the formal language and the structure it is interpreted over: the resulting theory can be axiomatized completely (in a second-order sense), and moreover is *categorical*: all models of this theory are isomorphic to the standard interpretation on the reals. This sort of axiomatization in very powerful logical

languages has found many successors, e.g., in qualitative axiomatizations of physics.

For our third example of a spatial logic, we turn to topology. While Euclidean geometry is associated with rigid transformations like translations, rotations, and inversions, the mathematicians creating topology in the early decades of the 20th Century focused on much coarser transformations deforming shapes up to tearing and knotting. Subsequent to its invention, topology, too, became an object of logical study, and yet again, Tarski's work proved seminal. Tarski observed that topology has small decidable fragments which could be brought to light by treating the topological interior operation as a *modal operator* (McKinsey and Tarski, 1944). The connection to the other spatial logics discussed above becomes apparent if we subject McKinsey and Tarski's original modal language to some essentially cosmetic reformulation. The variables of this language are taken to range over arbitrary subsets of any fixed topological space. These variables may be combined to form complex terms by means of function-symbols denoting various set-theoretic operations (union, intersection and complement), and topological operations (interior and closure); such terms denote subsets of the topological space over which they are interpreted. With terms constructed in this way, the language then features equality as its only predicate. Here we have extreme poverty of expressive resources: primitive function-symbols expressing only set-theoretic and topological operations, no non-logical predicates, and no quantifiers. But there is again a computational reward: the satisfiability problem for this logic is decidable in polynomial space. While too inexpressive to represent much of topology, this language has had profound repercussions in other areas, in particular in the universal algebra of Boolean algebras with added operators, and much contemporary modal logic. Furthermore, it has also been the inspiration for much recent work on topological logics, many of them equipped with more elaborate syntax and richer topological primitives, as the reader of this book will soon discover.

With these examples to guide us, let us return to the abstract characterization of spatial logic with which we began. Spatial logics arise by making a number of design choices, along three principal dimensions. The first concerns the collection of geometrical entities which make up our interpretations: points, lines, regions (of various kinds), and so on. In Tarski's (plane) Elementary Geometry, variables range over the collection of points in the Euclidean plane; likewise, in his Geometry of Solids, variables range over the collection of regular closed subsets of $\mathbb{R}^3$; and in his modal topological language, variables range over the collection of all subsets of some topological space. The second principal dimension concerns the choice of primitive relations and operations over these entities to interpret the non-logical primitives of our language. This choice of primitives of course reflects the level of spatial structure the particular logic is concerned with—metric, affine, projective, or topological; but even within

these broad divisions, there is room for almost endless variation. The third principal dimension concerns the purely logical resources at our disposal. We have already seen that these can be set at many levels: from weak 'constraint' languages through to richer first-order languages or even higher-order formalisms which include the resources of set theory. Needless to say, none of the choices along these principal dimensions is intrinsically right or wrong: they simply parametrize the family of available spatial logics.

Classification of geometrical languages in terms of the range of the spatial primitives they feature of course recalls the long-standing classification of 'geometries', broadly conceived, given by Klein's *Erlanger Programm* (Klein, 1893b; see also Klein, 1893a). And indeed the most sophisticated accounts of expressive power of such languages today are couched in terms of invariance relations between models (isomorphism, bisimulation, and the like), much in the same spirit. However, the logical approach opens up many new possibilities in this regard, such as, for instance, a new sort of invariance between topological patterns, much coarser even than topological homeomorphism, viz. modal bisimulation. This is topology taken to the extreme, but there are interesting interpretations in terms of model comparison games—a style of thinking which might have appealed to the founders of geometry, given Brouwer's early use of games in defining the notion of topological dimension (Brouwer, 1913, p. 148).

Once we have fixed a spatial logic, four salient issues present themselves. First, how can we characterize its valid formulas? Second, what is its expressive power? Third, what is its computational complexity? And fourth, what alternative interpretations does it have? We briefly consider each of these in turn. The first issue is so familiar as to require little explanation. Given a formal language interpreted over a certain class of geometrical structures, it typically makes sense to ask (depending on details of syntax) which sentences of that language are true in all structures of that class. Mostly, these characterizations are couched in the form of a list of axioms and (finitary) rules-of-inference. However, there are cases where additional machinery is required, for example, where the set of validities is not recursively enumerable, or where explicit proof systems are required to provide geometrical 'constructions' in Euclid's sense.

Second, we have already noted that current treatments of expressive power in logic are derived from the geometrical notions of invariance relations across models, setting the level of semantic resolution beyond which the given language cannot probe. Examples of such invariance relations are potential isomorphism for first-order logic, or bisimulation for modal languages; but there are many more. Within given models, such relations specialize to notions of automorphism or internal bisimulation—a viewpoint which is actually somewhat closer to the mathematician's usual way of thinking about 'symmetries' of a spatial structure. Weyl at one point observed that point tuples in Euclidean space which are related by an automorphism must satisfy the same geometrical

formulas, and raised the converse question of whether sharing the same properties in some given logical language implies automorphism invariance (Weyl, 1949, p. 73). Indeed, invariance is not just descriptive weakness, but also the source of information flow across situations! Logical model theory has a host of sophisticated results concerning invariance. In particular, invariance relations can be fine-tuned in terms of games, such as Ehrenfeucht-Fraïssé games matching first-order logic. Given a notion of invariance, the model theory of definability can start, and indeed, many results about expressive power of spatial languages can be found in the chapters to follow.

Third, complexity-theoretic analyses of logical systems typically focus on two problems: model-checking (determining whether a given formula is true in a *given* interpretation) and satisfiability checking (determining whether a given formula is true in *some* interpretation *or other*). Model-checking has been little-explored in the context of spatial logics; satisfiability checking, by contrast, has received much more attention. Most first- (or higher-) order spatial logics interpreted over familiar spatial domains are undecidable; therefore, this issue is obviously of greatest interest when dealing with spatial logics with more limited expressive power. A striking example is provided by spatial logics interpreted over the regular closed sets of arbitrary topological spaces whose language involves just Boolean connectives (no quantifiers) and whose spatial primitives represent various topological relations and functions. The satisfiability problem for such logics is generally decidable, and its complexity has been determined for a range of cases. In this light, spatial logics actually do pose an interesting challenge which is not yet well-understood. The general methodology in logic design has been to find expressive yet decidable formalisms, cleverly steering a middle course between the opposing evils of expressive poverty and undecidability. However, methods of analysis which work with general models are often powerless when confronted with languages interpreted over specific structures, as is generally the case with spatial logics. Sometimes, the spatial models over which one is working themselves support decidability for rich languages—witness again Elementary Geometry, where it is the structure of Euclidean space that drives the quantifier elimination procedure establishing decidability. We are still far from understanding the precise balance between all these triggers of higher or lower complexity in spatial logics.

Fourth, and most speculatively, we have the issue of alternative interpretations. Tarski's Geometry of Solids possesses, as we have seen, just one model up to isomorphism, but most spatial logics have many models. To some extent, this is just the expression of a familiar phenomenon in logic, and mathematics generally. Some theories, such as group theory or the theory of affine spaces, are designed to have many models, and the more of these there are, the greater their range of applicability. Other theories were intended to describe one particular structure, such as the natural numbers, Euclidean space, or most

imperialistically of all, the set-theoretic universe. Geometry provided early examples of how theories originally conceived as characterizations of specific structures could turn out to have alternative models. This issue is brought to the fore in the subject of spatial logic, where the formal systems under investigation expressly invite the search for alternative interpretations and thus alternative ways of conceptualizing space. Even bolder views were ventured by Beth in the 1950s, who claimed that it was geometry's move from one unique Space to a plurality of 'spaces' that underlay the system-based methodology of modern science and the fall of Aristotelian *a priori* dogmatism (Beth, 1959, Sec. 21). Be that as it may, the present editors agree that spatial logic can have philosophical repercussions beyond its narrower technical confines.

More prosaically, much of the renewed interest in spatial logic in recent years has come from computer science. We identify three examples of this trend. The first comes from artificial intelligence, where attempts have recently been made to develop logics of *qualitative spatial reasoning*. The motivation is as follows: numerical co-ordinate descriptions of the objects which surround us are hard to acquire, inherently error-prone, and probably unnecessary for most everyday tasks we want to perform (or want a machine to perform); therefore—so goes the argument—reasoning with purely qualitative descriptions of those objects' spatial configurations is closer to human reasoning and thus will lead to more efficient and effective AI. But *which* qualitative spatial terms, exactly, should we reason with? Ready-made tools from geometry or topology will not do: we have to devise new logics for ourselves. Many of these logics are discussed in this book.

The second example comes from the theory of spatial databases. In computer applications, spatial data is frequently stored in the form of polygons (or polyhedra)—in effect, sets of points definable by Boolean combinations of linear inequalities. These sets can be finitely represented, and their well-behaved character makes them particularly amenable to computer processing. But in fact there is no need to set our expressive sights so low; for polygons and polyhedra are a special case of the more general class of *semi-algebraic* sets, that is, those sets of points definable by Boolean combinations of polynomial inequalities. Within mathematics, semi-algebraic sets form the basis for real algebraic geometry; within computer science, they have given rise to the discipline of *constraint databases*. In a constraint database, spatial data is stored in the form of first-order formulas in the language of fields. The key fact here is the quantifier-elimination theorem for the theory of the reals. This result allows constraint databases to be accessed effectively using queries which are likewise written as first-order formulas over an appropriate vocabulary. The relevant chapter in this book explores some of the intricate logical issues that arise from this approach to spatial data.

Our third example comes from image processing, where it is convenient to describe objects as sets of vectors that can be 'added' (taking all linear sums) or 'subtracted' (taking all linear differences). By variously combining these 'Minkowski operations', certain useful processing tasks can be performed, as, for example, when one set of vectors, representing an 'eraser', is used to 'clean up' the boundary of another, representing a perceived object. *Mathematical morphology* is a theory of subsets of vector spaces with the two operations of addition and subtraction at its core; the properties of these operations are generalized in abstract algebraic and category-theoretic ways. Looking at space in this way brings to light a surprising amount of new structure. This theory was not developed within mathematical logic; but the relevant chapter in this book will show how logical patterns do arise, involving both modal and first-order languages, while the calculus of valid principles shows surprising analogies with logical systems proposed in recent decades for very different purposes, such as linear logics of computational resource management. Again, we see how new choices of spatial objects and spatial structures lead to new mathematics—and there is no reason to think that this creative process has yet run dry.

Finally, let us remove a possible misunderstanding, again taking a cue from the history of geometry. Our presentation may have made it look as if there is a vast collection of different spatial logics, each a world unto itself in terms of objects, primitive relations, and logical strength. But one of the most striking discoveries in the foundations of geometry in the 19th Century, prominently displayed in Hilbert's *Grundlagen*, was the fact that very different-looking theories can turn out to be related at a deeper level of analysis. Inspiring examples are the embeddings of non-Euclidean spaces into Euclidean ones given by Klein and Poincaré. Likewise, spatial logics show inter-connections which may be brought out by various means: semantic model transformations, direct linguistic translations, and so on. Even though little is known about the precise links between most known systems, we emphasize this point as a reassuring thought about the coherence of the field.

This concludes the editors' thoughts about the general setting for this book, while providing a way of positioning specific chapters. But of course, the real content is in the chapters themselves, which do much more than fit editorial preconceptions. Each tells a story about a particular approach to spatial logic. The chapters have been arranged in the following thread, though they can be read in other orders as well.

We start in Tarski's geometrical spirit, with first-order languages. In Ch. 2, Pratt-Hartmann considers first-order topological languages interpreted over low-dimensional Euclidean spaces, applying techniques from logical model theory to analyze expressive power and axiomatizability. In Ch. 3, Bennett and Düntsch study both first-order and weaker modal topological languages over a large class of topological spaces, emphasizing basic decidable

structures of wide use in AI and beyond. Renz and Nebel take this even further in Ch. 4, with syntactically highly restricted constraint languages for spatial structures, allowing for great computational efficiency.

From fragments of first-order languages, there is a natural transition to modal logics for topology, continuing the tradition started by Tarski and others in the 1930s. Ch. 5 by van Benthem and Bezhanishvili tells the story of modern modal approaches to topology (and a few other spatial structures), emphasizing the main axiomatic and semantic techniques developed in modern modal logic. This theme is then continued in Ch. 6 by Parikh, Moss and Steinsvold, who explore the other logical tradition of thinking about topology, viz. as an account of information structure. Next, Ch. 7 by Balbiani, Goranko, Kellerman and Vakarelov takes the modal viewpoint to the study of affine and metric geometry, moving up to first-order languages where needed. In particular, completeness theorems turn out to be related to the basic geometrical issue of coordinatization. Finally, Ch. 8 by Vickers takes the epistemic view of topology to the higher mathematical level of topos theory, merging spatial logic and epistemic logic with category theory and type theory.

Just as in science generally, so too in spatial logic, space enters into natural combinations with other fundamental notions. One obvious case is the combination of space and time, which is unavoidable in many practical computational settings, and of course, also, in the foundations of physics. Ch. 9 by Kontchakov, Kurucz, Wolter and Zakharyaschev studies temporal logics with added affine and metric modalities, using sophisticated techniques from current research on the complexity of combined modal logics. A special case of this type of combination is found in Ch. 10 by Kremer and Mints, who add a dynamic temporal operator of one-step system evolution to modal logics of topology, and show that this simple move provides significant results like the Poincaré recurrence theorem. Finally, Ch. 11 by Andréka, Madarász and Németi goes far beyond simple modal languages of space-time, and develops both the special and the general theory of relativity on a first-order basis, continuing Tarski's program for geometry to obtain striking new foundational results which are at the same time conceptually enlightening.

The next group of chapters represent a counterpoint to the 'logical' investigations so far, reporting further mathematical and computational advances. Ch. 12 by Smyth and Webster explores the extent to which topological ideas can be developed in discrete spaces, moving closer to the discrete topologies used in modern mathematics, pattern recognition, and image processing. Ch. 13 by Geerts and Kuijpers describes the use of algebraic constraints for spatial databases to describe regions in Euclidean space, reminding us of the great tradition of analytic geometry which also underlies the coordinatizations employed by Tarski, and by several authors in our book. Ch. 14 by Bloch,

Heijmans and Ronse develops the theory of mathematical morphology, both on concrete vector spaces and in algebraic abstraction, and introducing, at the end, logical formalisms based on them.

Beyond these technical subjects, our book still has a coda. We have indicated already that spatial logic also has a broader conceptual aspect. Ch. 15 by Varzi is an extensive discussion of spatial structure in the philosophical tradition, both ancient and modern, using logical tools to develop philosophical conceptions.

Despite the wealth of topics in our fifteen chapters, this book also set itself definite limits. First, we have not even exhausted the mathematical connection, witness the long-standing historical interest in 'diagrammatic reasoning' spawned by Euclid's Elements, and reinforced by modern research on graphical representation of information and associated styles of inference. There are deep issues here about the connection between symbolic and visual paradigms, bypassed in our cheerfully technical account of 'spatial logics'. We acknowledge them; but they are beyond the scope of this book. Likewise, many further varieties of spatial representation and spatial reasoning occur in disciplines like linguistics and psychology, and many more patterns await formal logical study. In addition, cognitive neuro-science tells us about the often surprising interplay between visual, diagrammatic, and more symbolically oriented parts of the brain in any reasoning task. Again, we think this is a fascinating theme, and we trust that many interesting interactions with the spatial logics of this book will one day come to light. But we have chosen the current set of chapters for their coherence in topic and methodology, and frankly also, their mathematical quality. We see the broader area of spatial reasoning; we recognize its relevance to the contents of this book; and exclusion does not imply disrespect. Broader texts on spatial reasoning should, and no doubt will, appear. But, in putting together this tighter book, the editors have stuck to what they see as the basic axiom of 'social geometry': *Always leave room for others*.

Acknowledgment

The authors wish to thank Dr. Paul van Ulsen for his kind help.

References

Beth, W. (1959). *The Foundations of Mathematics: a Study in the Philosophy of Science*. North Holland, Amsterdam.

Brouwer, L.E.J. (1913). Über den natürlichen Dimensionsbegriff. *Journal für die reine und angewandte Mathematik*, 142:146–152.

Hilbert, D. (1909). *Grundlagen der Geometrie*. B.G. Teubner, Leipzig and Berlin, 3rd edition.

Hilbert, D. (1950). *The Foundations of Geometry*. Open Court, La Salle, IL, 2nd edition.

Klein, F. (1893a). A comparative review of recent researches in geometry. *Bulletin of the New York Mathematical Society*, 2:215–249.

Klein, F. (1893b). Vergleichende Betrachtungen über neuere geometrische Forschungen. *Mathematische Annalen*, 43(1):63–100.

McKinsey, J. and Tarski, A. (1944). The algebra of topology. *Annals of Mathematics*, 45:141–191.

Tarski, A. (1956). Foundations of the geometry of solids. In *Logic, Semantics, and Metamathematics*, pages 24–29. Clarendon Press, Oxford.

Tarski, A. (1959). What is Elementary Geometry? In Henkin, L., Suppes, P., and Tarski, A., editors, *The Axiomatic Method, with Special Reference to Geometry and Physics*, pages 16–29. North-Holland Publishing Co., Amsterdam.

Weyl, H. (1949). *Philosophy of Mathematics and Natural Science*. Princeton University Press, Princeton, NJ.

Chapter 2

FIRST-ORDER MEREOTOPOLOGY

Ian Pratt-Hartmann
University of Manchester

Second Reader

Ivo Düntsch
Brock University

1. Introduction

One of the many achievements of coordinate geometry has been to provide a conceptually elegant and unifying account of the nature of geometrical entities. According to this account, the one primitive spatial entity is the point, and the one primitive geometrical property of points is coordinate position. All other geometrical entities—lines, curves, surfaces and bodies—are nothing but collections of points; and all properties and relations involving these entities may be defined in terms of the relative positions of the points which make them up. The success and power of this reduction is so great that the identification of spatial regions with the sets of points they contain has come to seem virtually axiomatic.

Over the years, however, various authors have expressed disquiet with this conceptual régime. The primary source of the disquiet is the conviction that our theory of space should use only those resources absolutely necessary to systematize the data of spatial experience. For *points* are such remote abstractions from the objects with which we daily interact, and *coordinate position* such a distant relative of the spatial properties and relations which we directly perceive, that the question arises as to whether alternative mathematical models of space are not possible—in particular, models in which the primitive spatial entities are not points, but regions, and in which the primitive spatial properties and relations are qualitative rather than quantitative.

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 13–97.
© 2007 *Springer.*

An example will help to make these worries more concrete. Consider any stable, medium-sized physical object, for example, a coffee cup. We all agree that this cup has a particular shape, which we may take to correspond to the region of space which it occupies at some instant. On the familiar point-based model of space, this region is a set of points. But suppose we now ask: is this set topologically open, semi-open or closed? That is: does it include none, some, or all of its boundary points? It is hard to see how we could answer this question. Not by microscopic analysis, since physical objects lose their definition on very small scales. And not by mathematical argument, since a world in which—say—cups are closed and saucers open is surely as logically possible as one where these topological characteristics are reversed. But if space really is made up of points as (modern) textbooks tell us, any assignment of a region of space to the coffee cup must answer the question. Perhaps then this model postulates too much.

This chapter addresses the question: what region-based accounts of the topological structure of space are possible? What can we say about them? How do they relate to each other and to the point-based models with which we are so familiar?

2. Mereotopologies

The purpose of this section is to outline the conceptual framework for region-based theories of space adopted in this chapter. Specifically, we introduce the concept of a *mereotopology* over a topological space, we discuss the role of mereotopologies as interpretations of signatures of topological primitives, and we list some key mathematical questions concerning them.

We assume familiarity with fundamental concepts and standard facts of point-set topology and Boolean algebra: for details, see, e.g. Kelley, 1955 and Koppelberg, 1989, Ch. 1, respectively. In the context of point-set topology, if u is any subset of a topological space X, we denote the interior of u by u^0 and the closure of u by u^-. (The more usual notations of $\bar{u}$ and $[u]$ for the closure of u are reserved for other purposes.) We write $\mathcal{F}(u)$ to denote the *frontier of* u, namely $u^- \setminus u^0$.

2.1 Regular open sets

How might we go about building a region-based model of the space we inhabit? The example of the coffee cup suggests that any such model should resolve the issue of frontier points. The following technical details are well-suited to this purpose.

DEFINITION 2.1 *Let u be a subset of some topological space X. We say that u is* regular open (in X) *if u is equal to the interior of its closure. We denote the set of regular open subsets of X by* $\mathrm{RO}(X)$.

To fix our intuitions, consider the space $X = \mathbb{R}^2$. The elements of $\mathrm{RO}(\mathbb{R}^2)$ are the open subsets of $\mathbb{R}^2$ having no "cracks" or "pin-holes" (Fig. 2.1). Corresponding remarks apply to the case $X = \mathbb{R}^3$. Taking regions of space to be

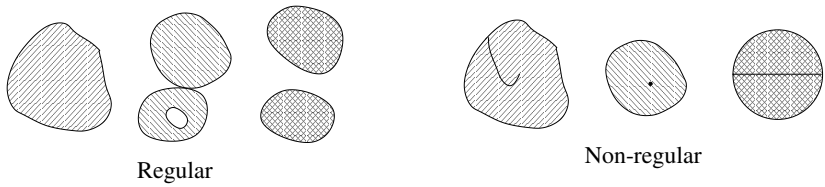


Figure 2.1. Some regular and non-regular open sets of the Euclidean plane.

regular open subsets of $\mathbb{R}^3$ finesses the issues encountered above concerning frontier points: regions are open by fiat. At the same time, however, it provides us with satisfying formal reconstructions of the intuitive notions of intersecting, merging and complementing regions, by means of the following standard theorem (see, for example, Koppelberg, 1989, pp. 25–27).

PROPOSITION 2.2 *Let X be a topological space. Then $\mathrm{RO}(X)$ is a Boolean algebra under the order $\subseteq$. In this Boolean algebra, top and bottom are defined by $1 = X$ and $0 = \emptyset$, and Boolean operations are defined by $x \cdot y = x \cap y$, $x + y = (x \cup y)^{-0}$ and $-x = X \setminus x^-$.*

Again, we can fix our intuitions regarding Proposition 2.2 by considering the case $X = \mathbb{R}^2$. The product, $x.y$, of two regular open sets x and y is simply their intersection, which is guaranteed to be a regular open set. The sum, $x+y$, of two regular open sets x and y is a little more complicated; *very* roughly, it is the union of x and y with any internal boundaries removed (Fig. 2.2). Finally, the complement, $-x$, of a regular open set x in $\mathrm{RO}(\mathbb{R}^2)$ is simply that part of the plane not occupied by x or its frontier. Corresponding remarks apply to the case $X = \mathbb{R}^3$.

It sometimes helps to reformulate the definition of regular open sets as follows. If $u \subseteq X$, then $\bigcup\{o \subseteq X | o \text{ open}, o \cap u = \emptyset\}$ is the largest open subset of X disjoint from u. We call this set the *pseudo-complement of* u, denoted u^*. From the above definitions, $u^* = X \setminus u^-$ and $u^{**} = (u^-)^0$. Hence, u is regular open if and only if $u = u^{**}$; and, if u is regular open, u^* is simply $-u$. The following lemma shows that every subset of X is 'close' to a regular open subset.

LEMMA 2.3 *Let X be a topological space. For every $u \subseteq X$, the set $r = (u^-)^0$ is an element of $\mathrm{RO}(X)$ such that $u^0 \subseteq r \subseteq u^-$. If u is open, then r is unique.*

Proof Obviously $u^0 \subseteq r \subseteq u^-$. To show that $(u^-)^0 \in \mathrm{RO}(X)$, it suffices to show that $u^{****} = u^{**}$. If v is any set at all, then $v^{**} \cap v^* = \emptyset$, whence $v^* \subseteq v^{***}$. Moreover, if o is any open set, then o^{***} is an open set disjoint from o^{**} and hence disjoint from every open set disjoint from o^* and hence disjoint from o itself, whence $o^{***} \subseteq o^*$. Thus, for any open set o, $o^{***} = o^*$. Since u^* is open, we have $u^{****} = u^{**}$. For the final statement, if $s \in \mathrm{RO}(X)$ also satisfies $u \subseteq s \subseteq u^-$, then the (regular) open sets $s \cdot -r$ and $r \cdot -s$ are both in $u^- \setminus u$ and so are empty. QED

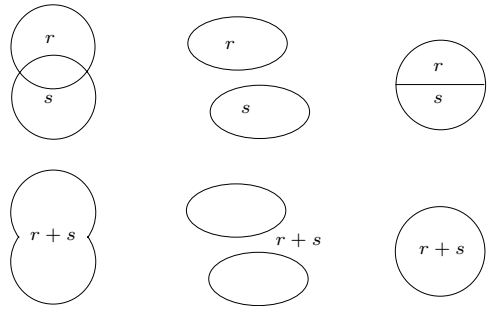


Figure 2.2. Three pairs of regions and (below) their sums in $\mathrm{RO}(\mathbb{R}^2)$.

For the above reasons, it has become common practice in discussions of region-based theories of space to model regions of space as regular open subsets of $\mathbb{R}^3$; and that is the approach we shall take here. In the sequel, we shall always use the letters r, s, t to range over regular open sets; when we are concerned only with regular open sets, we write $r \leq s$ in preference to $r \subseteq s$, 0 in preference to $\emptyset$ and $r \cdot s$ in preference to $r \cap s$. Resorting to regular open sets is of course not the only way of dealing with boundary disputes. One obvious alternative is to use regular closed sets (sets equal to the closures of their interiors), since the regular closed sets of any topological space also form a Boolean algebra, which is in fact isomorphic to the Boolean algebra of regular open sets. Thus, in modelling regions as regular open sets of $\mathbb{R}^3$, it is understood that it is the resulting structure that is important, not the precise constitution of its elements. Understanding what this idea means in detail forms a central theme of this chapter.

We conclude our discussion of regular open sets by proving some technical results which will be useful below. Recall in this context that, if u, v are connected subsets of a topological space, with $u \cap v \neq \emptyset$, then $u \cup v$ is connected. Moreover, if u is connected and $u \subseteq v \subseteq u^-$, then v is connected.

LEMMA 2.4 *Let X be a topological space, let $u, v \subseteq X$ and let $r, s \in \mathrm{RO}(X)$. We have:*

(i) $(u \cup v)^{-0} = u^{-0} + v^{-0}$;

(ii) $r \cup s \subseteq r + s \subseteq r \cup s \cup (r^- \cap s^-) \subseteq (r \cup s)^-$;

(iii) $(r + s)^- = r^- \cup s^- = (r \cup s)^-$;

(iv) *if* r *and* s *are connected with* $r \cdot s > 0$, *then* $r + s$ *is connected.*

Proof (i) By Lemma 2.3, $(u \cup v)^{-0}$ is a regular open set which evidently contains the regular open sets u^{-0} and v^{-0}. Certainly, then $u^{-0} + v^{-0} \subseteq (u \cup v)^{-0}$. For the reverse inclusion, $(X \setminus u)^0 \cap (X \setminus v)^0 \cap (u \cup v)^{-0} = \emptyset$, whence $(X \setminus u)^{0^{-0}} \cap (X \setminus v)^0 \cap (u \cup v)^{-0} = \emptyset$, whence $(X \setminus u)^{0^{-0}} \cap (X \setminus v)^{0^{-0}} \cap (u \cup v)^{-0} = \emptyset$, whence $((X \setminus u)^{0^-} \cap (X \setminus v)^{0^-})^0 \cap (u \cup v)^{-0} = \emptyset$, whence $((X \setminus u)^{0^-} \cap (X \setminus v)^{0^-})^{0^-} \cap (u \cup v)^{-0} = \emptyset$. That is: $(u \cup v)^{-0} \subseteq (u^{-0} \cup v^{-0})^{-0}$. But by Proposition 2.2, $(u^{-0} \cup v^{-0})^{-0} = u^{-0} + v^{-0}$.

(ii) The only non-trivial inclusion is $r + s \subseteq r \cup s \cup (r^- \cap s^-)$. So suppose $p \notin s$ and $p \notin r^-$. That is, $p \in (-s)^-$ and $p \in -r$. But then, for all open o with $p \in o$, $o \cap -r$ is also open with $p \in o \cap -r$, whence $(o \cap -r) \cap -s \neq \emptyset$—that is, $o \cap (-r \cdot -s) \neq \emptyset$. Hence $p \in (-r \cdot -s)^-$ so $p \notin -(-r \cdot -s) = r + s$. A similar argument applies if $p \notin r$ and $p \notin s^-$.

(iii) $(r+s)^- = X \setminus -(r+s) = X \setminus (-r \cdot -s) = (X \setminus -r) \cup (X \setminus -s) = r^- \cup s^-$.

(iv) Certainly, $r \cup s$ is connected, and by (ii), $r \cup s \subseteq r + s \subseteq (r \cup s)^-$, whence $r + s$ is connected. QED

We note in passing that determining the validity of statements such as those of Lemma 2.4 is actually a decidable problem. See, e.g. Cantone and Cutello, 1994, Nutt, 1999, Pratt-Hartmann, 2002 and, for a fuller discussion, Ch. 9.

2.2 Mereotopologies

We have argued, provisionally, that, for a subset of $\mathbb{R}^3$ to count as a region, it should be regular open. But it would be hasty to assume that *all* regular open subsets of $\mathbb{R}^3$ should count as regions, at least if regions are supposed to be parts of space occupied (or left unoccupied) by physical objects, since $\mathrm{RO}(\mathbb{R}^3)$ contains some pathological members. This presents us with the question: if not all regular open subsets of $\mathbb{R}^3$ qualify as *bona fide* regions, which do? As we shall see, the answers available and the issues which hinge on them require detailed analysis.

In view of these uncertainties, we adopt a very general notion of a region-based model of space—just sufficiently constrained that we can sensibly confine attention to the structure of regions in question without worrying about the points

of which they are composed. In the context of point-set topology, a topological space is commonly said to be *semi-regular* if it has a basis of regular open sets, and *locally connected* if it has a basis of connected sets. It easy to see that, in a locally connected space, every component of an open set is open. Recall also, in the context of Boolean algebras, that, if B is a Boolean algebra and B' a Boolean sub-algebra of B, then B' is said to be *dense* (*in* B) if, for every $b \in B$ with $0 < b$, there exists $b' \in B'$ with $0 < b' \leq b$.

DEFINITION 2.5 *Let X be a topological space. A* mereotopology over X *is a Boolean sub-algebra M of* $\mathrm{RO}(X)$ *such that, if o is an open subset of X and $p \in o$, there exists $r \in M$ such that $p \in r \subseteq o$. We refer to the elements of M as* regions. *If M is a mereotopology such that any component of a region in M is also a region in M, then we say that M* respects components.

Note that a mereotopology over X is always a dense sub-algebra of $\mathrm{RO}(X)$. Our first task is to check that $\mathrm{RO}(X)$ *is* a mereotopology, for a suitable class of topological spaces.

LEMMA 2.6 *Let X be a semi-regular space. Then* $\mathrm{RO}(X)$ *is a mereotopology over $\mathbb{R}^n$; if X is also locally connected, then* $\mathrm{RO}(X)$ *respects components.*

Proof The first part of the lemma is instant from the relevant definitions. For the second part, let $r \in \mathrm{RO}(X)$, and let s be a component of r. Since X is locally connected, s is open, whence, by Lemma 2.3, $(s^-)^0$ is regular open with $s \subseteq (s^-)^0 \subseteq s^-$. Then, s^{-0} is a connected subset of r including s, whence $s^{-0} = s$ by the maximality of s. QED

Some etymological explanation is in order here. The term *mereology* was first introduced by Leśniewski, and denotes the logic of the part-whole relationship. (For a survey, see, e.g. Simons, 1987.) The term *mereotopology* is a much more recent coinage, and standardly denotes the study of topological relationships in which regions, rather than points, are the primitive objects. (It is unclear where the word first appeared in print.) The employment of the word as a count-noun in Definition 2.5, to denote a certain class of mathematical structures, is new here, and prompted by analogy with the parallel usage of the word *topology*.

The foregoing discussion suggests that our search for a region-based model of space should begin with an examination of mereotopologies over $\mathbb{R}^3$. This approach may at first seem dissatisfying, because it depends for its formulation on the very point-based model of space we are trying to escape. As we shall see, however, it is the *structure* of the resulting collection of regions that will interest us—and the characterization of that structure in purely intrinsic terms forms one of the main themes of this chapter. But before we can seek such intrinsic characterizations, we must first clarify what it is we want to characterize.

2.3 Geometric mereotopologies

The question before us is to identify the regular open subsets of $\mathbb{R}^3$ which we are prepared to count as "sensible" regions of space. Here is a standard answer from the mathematical literature. Let L' be the first-order language with the arithmetic signature $(<, +, \cdot, 0, 1)$, interpreted over $\mathbb{R}$ in the usual way. (This interpretation is of course completely separate from our use of the same symbols to denote Boolean operations on regular open sets!) For the purposes of this chapter, we may say that a set $u \subseteq \mathbb{R}^n$ is *semi-algebraic* if there exists an L'-formula $\phi(\bar{x}, \bar{y})$ in $n + m$ variables $\bar{x}, \bar{y}$ and an m-tuple of real numbers $\bar{b}$ such that

$$u = \{\bar{a} \in \mathbb{R}^n | \text{ the } (n+m)\text{-tuple } \bar{a}, \bar{b} \text{ satisfies the formula } \phi(\bar{x}, \bar{y})\}.$$

For a detailed discussion of semi-algebraic sets, see, e.g. van den Dries, 1998, Bochnak et al., 1998 and also Ch. 13. (The more standard definition of semi-algebraic sets is equivalent to ours, and makes the name less puzzling.) For mereotopological purposes, we are exclusively interested in those semi-algebraic subsets of $\mathbb{R}^n$ which are *regular open*.

DEFINITION 2.7 *For $n > 0$, we denote the set of regular open, semi-algebraic sets in $\mathbb{R}^n$ by* $\mathrm{ROS}(\mathbb{R}^n)$.

LEMMA 2.8 *For $n > 0$,* $\mathrm{ROS}(\mathbb{R}^n)$ *is a mereotopology over $\mathbb{R}^n$.*

Proof We first show that $\mathrm{ROS}(\mathbb{R}^n)$ is a Boolean sub-algebra of $\mathrm{RO}(\mathbb{R}^n)$. Evidently, $0, 1 \in \mathrm{ROS}(\mathbb{R}^n)$. Moreover, if a set u is definable by a first-order formula in the language of arithmetic, then so are its closure u^- and its interior u^0. Hence, if $r, s \in \mathrm{ROS}(\mathbb{R}^n)$, then so are $r \cdot s = r \cap s$, $r + s = (r \cup s)^{-0}$ and $-r = \mathbb{R}^n \setminus r^-$. We must establish that, for $p \in o$ with $o \subseteq \mathbb{R}^n$ open, there exists $r \in \mathrm{ROS}(\mathbb{R}^n)$ such that $p \in r \subseteq o$. But this is obvious since any open ball is an element of $\mathrm{ROS}(\mathbb{R}^n)$. QED

The structure of regular open semi-algebraic subsets of $\mathbb{R}^3$ might have a better claim to count as a region-based model of space than the whole of $\mathrm{RO}(\mathbb{R}^3)$, because it does a good job of ruling out various pathological regular open sets.

More generally, semi-algebraic sets count as well-behaved. One of their fundamental properties is that they admit of "cell decompositions". If $d > 0$, a *d-cell* in $\mathbb{R}^n$ is any semi-algebraic subset of $\mathbb{R}^n$ homeomorphic to the open d-dimensional ball; a 0-*cell* in $\mathbb{R}^n$ is a singleton; and a *cell* is a d-cell for some d ($0 \leq d \leq n$). The following result is standard (van den Dries, 1998, Ch. 3, Theorem 2.11).

PROPOSITION 2.9 (CELL DECOMPOSITION THEOREM) *If u is a semi-algebraic subset of $\mathbb{R}^n$, then u is the union of a finite collection of pairwise disjoint, semi-algebraic cells.*

For regular open semi-algebraic sets, this yields:

LEMMA 2.10 *Every* $r \in \mathrm{ROS}(\mathbb{R}^n)$ *is the sum of finitely many pairwise disjoint* n*-cells in* $\mathrm{ROS}(\mathbb{R}^n)$.

Proof By Proposition 2.9, let $r = u_1 \cup \cdots \cup u_m$ where the u_i are pairwise disjoint, semi-algebraic cells. Since r is regular, $r = r^{-0} = (u_1 \cup \cdots \cup u_m)^{-0} = u_1^{-0} + \cdots + u_m^{-0}$, by Lemma 2.4 (i). If u_i is a d-cell for $d < n$, then $u_i^{-0} = 0$; if u_i is an n-cell, $u_i^{-0} = u_i$. QED

The following notion will play an important part in the ensuing discussion.

DEFINITION 2.11 *A mereotopology* M *is* finitely decomposable *if every region in* M *is the sum of finitely many connected regions in* M.

LEMMA 2.12 $\mathrm{ROS}(\mathbb{R}^n)$ *is finitely decomposable.*

Proof By Lemma 2.10, since cells are connected. QED

LEMMA 2.13 *Every finitely decomposable mereotopology* M *over a locally connected space* X *respects components; moreover, every region in* M *is the sum of its components.*

Proof Suppose $r \in M$, and s is a component of r. By Lemma 2.6, $s \in \mathrm{RO}(X)$. Let $r_1, \ldots, r_n$ be connected elements of M such that $r = r_1 + \cdots + r_n$. By the maximality of s and Lemma 2.4 (iv), either $r_i \leq s$ or $r_i \cdot s = 0$ for all i $(1 \leq i \leq n)$. Thus, s is the sum of those r_i such that $r_i \leq s$. QED

Of course, the converse of Lemma 2.13 is false: although $\mathrm{RO}(X)$ respects components for any locally connected space X, it is easy to see that, for example, $\mathrm{RO}(\mathbb{R}^n)$ is not finitely decomposable for any $n > 0$.

The mereotopology $\mathrm{ROS}(\mathbb{R}^n)$ is thus at least a plausible region-based model of the space we inhabit. But it is not the only candidate for this job. Observe that any $(n-1)$-dimensional hyperplane of $\mathbb{R}^n$ cuts $\mathbb{R}^n$ into two residual domains, which we shall call *half-spaces*. It is easy to see that these half-spaces are regular open, with each being the pseudo-complement of the other. Hence, we can speak about the sums, products and complements of half-spaces in $\mathrm{RO}(\mathbb{R}^n)$.

DEFINITION 2.14 *A* basic polytope *in* $\mathbb{R}^n$ *is the product, in* $\mathrm{RO}(\mathbb{R}^n)$*, of finitely many half-spaces. A* polytope *in* $\mathbb{R}^n$ *is the sum, in* $\mathrm{RO}(\mathbb{R}^n)$*, of any finite set of basic polytopes. We denote the set of polytopes in* $\mathbb{R}^n$ *by* $\mathrm{ROP}(\mathbb{R}^n)$*; we call the polytopes in* $\mathrm{ROP}(\mathbb{R}^2)$ polygons *and those in* $\mathrm{ROP}(\mathbb{R}^3)$ polyhedra.

Thus, polytopes (in our sense) may be unbounded, disconnected, and may have disconnected complements. Fig. 2.3 shows a selection of polygons. (In

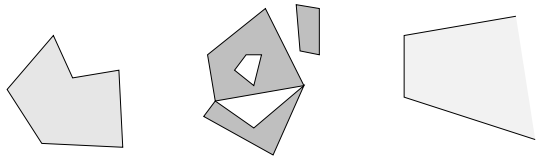


Figure 2.3. Three (differently shaded) regions in the mereotopology $\mathrm{ROP}(\mathbb{R}^2)$.

alternative parlance, the elements of $\mathrm{ROP}(\mathbb{R}^n)$ are the regular open *semi-linear* sets.) Evidently, the polyhedra constitute a more parsimonious region-based model of space than does $\mathrm{ROS}(\mathbb{R}^3)$.

Indeed, the following construction gives us a more parsimonious spatial ontology still. If an $(n-1)$-dimensional hyperplane in $\mathbb{R}^n$ is defined by an equation $a_0 + a_1x_1 + \cdots + a_nx_n = 0$, where the a_i $(0 \leq i \leq n)$, are rational numbers, we call it a *rational hyperplane*; and if a half-space is bounded by a rational hyperplane, we call it a *rational half-space*. Now we define:

DEFINITION 2.15 *A* basic rational polytope *in* $\mathbb{R}^n$ *is the product, in* $\mathrm{RO}(\mathbb{R}^n)$*, of finitely many rational half-spaces. A* rational polytope *in* $\mathbb{R}^n$ *is the sum, in* $\mathrm{RO}(\mathbb{R}^n)$*, of any finite set of basic rational polytopes. We denote the set of rational polytopes in* $\mathbb{R}^n$ *by* $\mathrm{ROQ}(\mathbb{R}^n)$*; we call the elements of* $\mathrm{ROQ}(\mathbb{R}^2)$ rational polygons *and those of* $\mathrm{ROQ}(\mathbb{R}^3)$ rational polyhedra.

Evidently, $\mathrm{ROQ}(\mathbb{R}^n) \subsetneq \mathrm{ROP}(\mathbb{R}^n) \subsetneq \mathrm{ROS}(\mathbb{R}^n) \subsetneq \mathrm{RO}(\mathbb{R}^n)$. Note that $\mathrm{ROQ}(\mathbb{R}^n)$ is countable.

LEMMA 2.16 *The collections* $\mathrm{ROP}(\mathbb{R}^n)$ *and* $\mathrm{ROQ}(\mathbb{R}^n)$ *are finitely decomposable mereotopologies over* $\mathbb{R}^n$*.*

Proof Basic polytopes are convex, and hence connected. QED

As models of the space we inhabit, $\mathrm{ROP}(\mathbb{R}^3)$ and $\mathrm{ROQ}(\mathbb{R}^3)$ may seem overly austere—for they contain no regions with curved boundaries. However, their study turns out to be instructive, as we shall see below.

2.4 Interpretations

So far, we have discussed various ways of selecting a collection of "regions" from among the subsets of $\mathbb{R}^n$. But this selection process only really becomes interesting when we consider formal languages whose variables range over these collections, and whose non-logical constants belong to a limited repertoire of spatial primitives.

We assume familiarity with basic first-order logic: for details, see Hodges, 1993, Ch 1. In this context, we employ the following standard notation and

terminology. Let Σ be a signature consisting of (zero or more) predicates, function-symbols and individual constants; we denote the first-order language with signature Σ by L_Σ. An L_Σ-formula with no free variables is called an L_Σ-*sentence*. Let $\mathfrak{A}$ be a structure interpreting the symbols in Σ over some domain A (assumed non-empty). For any L_Σ-formula $\phi(\bar{x})$, with $n > 0$ free-variables $\bar{x}$ and any n-tuple $\bar{a}$ from A, we write $\mathfrak{A} \models \phi[\bar{a}]$ if $\bar{a}$ satisfies $\phi(\bar{x})$ in $\mathfrak{A}$; similarly, for any L_Σ-sentence ϕ, we write $\mathfrak{A} \models \phi$ if ϕ is true in $\mathfrak{A}$. We call $\{\psi \mid \psi$ an L_Σ-sentence and $\mathfrak{A} \models \psi\}$ the L_Σ-*theory of* $\mathfrak{A}$, denoted $\mathrm{Th}_\Sigma(\mathfrak{A})$. Two structures $\mathfrak{A}$ and $\mathfrak{B}$ are *elementarily equivalent* (*for* Σ), written $\mathfrak{A} \equiv_\Sigma \mathfrak{B}$, if $\mathrm{Th}_\Sigma(\mathfrak{A}) = \mathrm{Th}_\Sigma(\mathfrak{B})$. We write $f : \mathfrak{A} \simeq_\Sigma \mathfrak{B}$ if f is a Σ-structure isomorphism from $\mathfrak{A}$ onto $\mathfrak{B}$ (and $\mathfrak{A} \simeq_\Sigma \mathfrak{B}$ if such an f exists). It is a simple result that if $f : \mathfrak{A} \simeq_\Sigma \mathfrak{B}$ and $\phi(\bar{x})$ is an L_Σ-sentence, then $\mathfrak{A} \models \phi[\bar{a}]$ implies $\mathfrak{B} \models \phi[f(\bar{a})]$ for every tuple $\bar{a}$ from A; in particular, $\mathfrak{A} \simeq_\Sigma \mathfrak{B}$ implies $\mathfrak{A} \equiv_\Sigma \mathfrak{B}$. We write $\mathfrak{A} \subseteq_\Sigma \mathfrak{B}$, if $\mathfrak{A}$ is a submodel of $\mathfrak{B}$ (i.e. $A \subseteq B$ and $\mathfrak{A}$ is the restriction of $\mathfrak{B}$ to A), and $\mathfrak{A} \preceq_\Sigma \mathfrak{B}$ if $\mathfrak{A}$ is an *elementary submodel* of $\mathfrak{B}$ (i.e. $A \subseteq B$ and and every tuple $\bar{a}$ of A satisfies the same L_Σ-formulas in both $\mathfrak{A}$ and $\mathfrak{B}$). We say that $\mathfrak{A}$ is *elementarily embeddable in* $\mathfrak{B}$ if $\mathfrak{A}$ is isomorphic to an elementary submodel of $\mathfrak{B}$. Trivially, $\mathfrak{A} \preceq_\Sigma \mathfrak{B}$ implies $\mathfrak{A} \equiv_\Sigma \mathfrak{B}$. Reference to the signature Σ, and the associated subscripts, is suppressed when clear from context.

Let M be a mereotopology over some topological space X. If Σ is a signature whose symbols conventionally denote familiar mereological or topological concepts, then M can always be regarded as a Σ-structure by interpreting the symbols of Σ in the familiar way. In particular, we take the symbols $0, 1, +, \cdot, -$ and $\leq$ to have the obvious (Boolean algebra) interpretations over M; similarly, we take the unary predicate c to denote the property of being connected, and the binary predicate C to denote the relation which holds between two regions if and only if their topological closures intersect. Table 2.1 gives a formal summary. Under these interpretations, we may regard any mereotopology M as an interpretation for the signature $\Sigma = (0, 1, +, \cdot, -, \leq, c, C)$, or any subset thereof. That is: any L_Σ-sentence has a truth-value in M, and any L_Σ-formula $\phi(\bar{x})$ with $n > 0$ free variables defines an n-ary relation over M, namely, the set of n-tuples from M satisfying $\phi(\bar{x})$. We remark that our interpretation of C is intended as a rational reconstruction of the relation which Whitehead, 1929 called "extensive connection", and which has historically played a prominent role in region-based theories of space. Since Whitehead's term risks confusion with the standard topological notion of *connectedness*, we follow more recent usage and read $C(x, y)$ as "x contacts y".

Some examples will help to clarify the issues that arise concerning first-order languages interpreted over mereotopologies.

EXAMPLE 2.17 *Let* $\Sigma = (C, c, \leq)$, *and let* ψ_{inf} *be the* L_Σ-*sentence*

$$\forall x \forall y (C(x, y) \rightarrow \exists z (c(z) \wedge z \leq y \wedge C(x, z))).$$

Symbol	Type	Interpretation
0	individual constant	$0^M = \emptyset$
1	individual constant	$1^M = X$
$+$	binary function	$+^M(r,s) = ((r \cup s)^-)^0$
$\cdot$	binary function	$\cdot^M(r,s) = r \cap s$
$-$	unary function	$-^M(r) = X \setminus r^-$
$\leq$	binary predicate	$\leq^M = \{\langle r,s\rangle \in M^2 \mid r \subseteq s\}$
c	unary predicate	$c^M = \{r \in M \mid r \text{ connected}\}$
C	binary predicate	$C^M = \{\langle r,s\rangle \in M^2 \mid r^- \cap s^- \neq \emptyset\}$

Table 2.1. Interpretations of common mereotopological primitives, where M is a mereotopology over a topological space X.

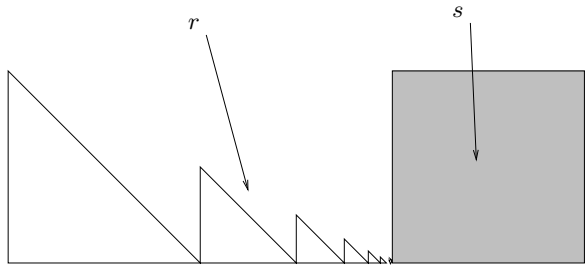


Figure 2.4. Two elements in $\mathrm{RO}(\mathbb{R}^2)$, one with infinitely many components.

This sentence "says" that, if a region contacts another region, then it contacts some connected part of it. Let M be any finitely decomposable mereotopology; then $M \models \psi_{\mathrm{inf}}$. For suppose $M \models C[r,s]$, and let $s_1, \ldots, s_m$, be connected regions of M summing to s. By Lemma 2.4(iii), $s^- = s_1^- \cup \cdots \cup s_m^-$, whence $M \models C[r, s_i]$ for some i. On the other hand, it is not difficult to see that $\mathrm{RO}(\mathbb{R}^2) \not\models \psi_{\mathrm{inf}}$. Fig. 2.4 shows two regular open regions r, s in the plane, where r has infinitely many components, and s touches the closure of r but is separated from each of its components.

Example 2.4 shows, in particular, that the differences between the region-based models of space $\mathrm{RO}(\mathbb{R}^3)$ and $\mathrm{ROS}(\mathbb{R}^3)$ are "visible" to certain first-order languages with signatures of topological primitives. In fact, the existence of regions with infinitely many components is not the only difference between these mereotopologies, as the next example shows.

EXAMPLE 2.18 *Let $\Sigma = (c, +)$, and let ψ_{sum} be the L_Σ-sentence*

$$\forall x_1 \forall x_2 \forall x_3 (c(x_1) \wedge c(x_2) \wedge c(x_3) \wedge c(x_1 + x_2 + x_3) \rightarrow (c(x_1 + x_2) \vee c(x_1 + x_3))).$$

This sentence "says" that if three connected regions have a connected sum, then the first must form a connected sum with one of the other two. We show

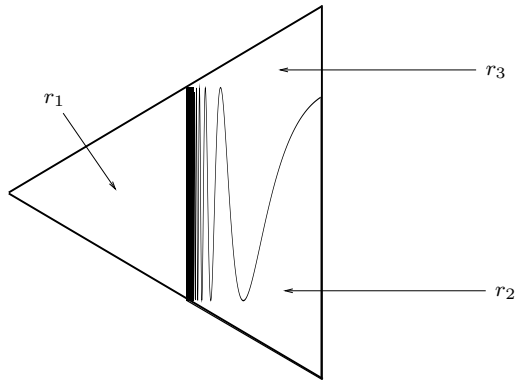


Figure 2.5. Three elements in $\mathrm{RO}(\mathbb{R}^2)$.

in Lemma 2.56 below that, if M is any of $\mathrm{ROS}(\mathbb{R}^2)$, $\mathrm{ROP}(\mathbb{R}^2)$ *or* $\mathrm{ROQ}(\mathbb{R}^2)$, *then $M \models \psi_{\mathrm{sum}}$. However, it turns out that* $\mathrm{RO}(\mathbb{R}^2) \not\models \psi_{\mathrm{sum}}$. *For let*

$$\begin{aligned} r_1 &= \{(x,y) | -1 < x < 0\,;\ -1-x < y < 1+x\} \\ r_2 &= \{(x,y) | 0 < x < 1\,;\ -1-x < y < \sin(1/x)\} \\ r_3 &= \{(x,y) | 0 < x < 1\,;\ \sin(1/x) < y < 1+x\}\,, \end{aligned}$$

as depicted in Fig. 2.5. It is easy to check that $r_1 + r_2 + r_3$ is the large triangle, and so is certainly connected, but that neither $r_1 + r_2$ nor $r_1 + r_3$ is connected.

We shall see in Sec. 5 that, in some sense, Examples 2.17 and 2.18 represent the *only* differences between $\mathrm{RO}(\mathbb{R}^2)$ and $\mathrm{ROS}(\mathbb{R}^2)$.

Our final example illustrates a rather different set of issues concerning first-order mereotopological theories. We require the following fact about the topology of Euclidean spaces (Newman, 1964, p. 137).

PROPOSITION 2.19 *If d_1 and d_2 are non-intersecting closed sets in $\mathbb{R}^n$, and points p and q are connected in $\mathbb{R}^n \setminus d_1$ and also in $\mathbb{R}^n \setminus d_2$, then p and q are connected in $\mathbb{R}^n \setminus (d_1 \cup d_2)$.*

EXAMPLE 2.20 *Let $\Sigma = (C, c, \cdot, -)$, and let ψ_{sep} be the L_Σ-sentence*

$$\forall x \forall y (c(x) \wedge c(y) \rightarrow (c(x \cdot y) \vee C(-x, -y))).$$

This sentence "says" that the closures of the complements of any two connected regions whose product is not connected intersect. Suppose that $r, s \in \mathrm{RO}(\mathbb{R}^n)$ are connected, with $r \cdot s$ not connected. Putting $d_1 = \mathbb{R}^n \setminus r$ and $d_2 = \mathbb{R}^n \setminus s$, we

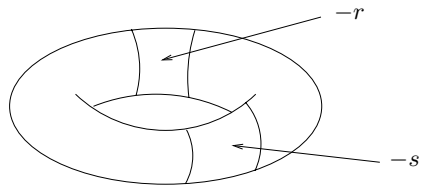


Figure 2.6. The (complements of) two connected elements r and s in the regular open algebra of a torus: $r \cdot s$ is not connected, and $(-r)^- \cap (-s)^- = \emptyset$.

have $d_1 \cup d_2 = \mathbb{R}^n \setminus (r \cdot s)$, whence, by Proposition 2.19, $(-r)^- \cap (-s)^- \neq \emptyset$. Thus, if M is a mereotopology over any of the spaces $\mathbb{R}^n$, $M \models \psi_{\text{sep}}$. However, ψ_{sep} is not true for all mereotopologies. For example, let X be a torus, let M be $\text{RO}(X)$, and let $r, s \in M$ be such that $-r$ and $-s$ are as illustrated in Fig. 2.6. By inspection, r and s are connected, $r \cdot s$ is not connected, and $-r$ does not contact $-s$. Hence, $M \models \neg\psi_{\text{sep}}$.

Thus the regular open algebras of the torus and the Euclidean plane have different first-order mereotopological theories over the signature $\{C, c, \cdot, -\}$.

There is nothing privileged about the above collection of primitives: in principle, we could employ any signature whose symbols can be given fixed interpretations over the structures we choose to confine our attention to. Since this chapter deals with *topological* notions, we consider only signatures with fixed *topological* interpretations—that is, signatures whose interpretations are preserved by homeomorphisms of the underlying topological space. For brevity, we speak of a 'signature of topological primitives'. For investigations of region-based theories with non-topological signatures, see, e.g. Davis et al., 1999, Pratt, 1999.

Given a mereotopology M and a signature Σ of topological primitives, three salient issues present themselves. The first concerns the *expressive power* of the first-order topological language L_Σ over the mereotopology M. Any L_Σ-formula $\phi(\bar{x})$ with free variables $\bar{x} = x_1, \ldots, x_n$ defines an n-ary relation over M—namely, the set of n-tuples $\bar{r}$ satisfying $\phi(\bar{x})$ in M. And it is therefore natural to ask *which* relations can be so defined, and in particular, which primitives can be defined in terms of which others. Of particular interest in this regard is the property of being topologically indistinguishable from a specific object or tuple of objects. That is, given a tuple $\bar{r}$ from M, we would particularly like to know whether L_Σ is expressive enough to give a topologically complete characterization of $\bar{r}$. The answers to these questions depend heavily on the mereotopology M: Sections 3 and 4 analyse the expressive power of various first-order topological languages for well-behaved mereotopologies over the

Euclidean plane. Sec. 6 analyses the much more difficult case of well-behaved mereotopologies over $\mathbb{R}^3$.

The second salient issue concerns the L_Σ*-theory* of M. Examples 2.17 and 2.18 show that restricting regions to be *semi-algebraic* (regular open) sets does affect the resulting first-order theory over some signatures of topological primitives. And the question therefore arises as to what other restrictions might be sensible, and what effect, if any, these restrictions have on the resulting first-order mereotopological theories. Most ambitiously, perhaps, we might ask whether the set of first-order sentences true in various mereotopologies can be axiomatically characterized. Sec. 5 provides an example of such an axiomatic characterization. As a by-product of this analysis, we show that a wide range of plane mereotopologies share the same L_Σ-theory for (most) topological signatures Σ, and we venture to take that theory as the *standard* first-order L_Σ-theory of plane mereotopology. In this sense, the choice of what, exactly, counts as a region is much less critical than we might at first have supposed.

The third salient issue concerns the ontological commitments entailed by first-order mereotopological theories. To understand this issue, recall that a mereotopology M is a collection of subsets of some topological space, which we have chosen to regard as a Σ-structure, for some signature Σ of topological primitives. Any such mereotopology M thus defines an L_Σ-theory $\mathrm{Th}_\Sigma(M)$. But of course, *any* Σ structure $\mathfrak{A}$ with $\mathrm{Th}_\Sigma(\mathfrak{A}) = \mathrm{Th}_\Sigma(M)$ can be thought of as a (region-based) model of space which, from the point of view of L_Σ, makes exactly the same predictions as M. It is therefore natural to ask *which* structures these are, and what, if anything, we can say about their relationship to M. Notice that the elements of such Σ-structures need not be regions of topological spaces at all; as such they are genuinely region-based theories of space. In particular, we may ask whether mereotopologies in general admit of intrinsic characterizations making no reference to the topological spaces whose regions they make up. And we may further ask—particularly in the light of Example 2.20—what information those intrinsic characterizations yield about the topological spaces in question. Sec. 7 answers these, and related, questions.

The above three issues constitute the primary agenda of mereotopology, as conceived here.

3. Defining topological relations

Our task in this section is to compare the relative expressiveness of first-order languages having different signatures of topological primitives. Our main result is that L_C is at least as expressive as $L_{c,\leq}$ over all sensible mereotopologies. We also show that over some mereotopologies of interest, $L_{c,\leq}$ is also at least as expressive as L_C.

We assume familiarity with the standard (T_i-) separation properties of topological spaces. Terminology varies here: we adopt the convention according to which T_i-separation for $i > 2$ does not by definition imply T_1-separation; and we say that a space X is *Hausdorff* if it satisfies T_2-separation, *regular* if it satisfies both T_3- and T_1-separation, and *normal* if it satisfies both T_4- and T_1-separation.) In addition, we occasionally employ the following less familiar separation property (Düntsch and Winter, 2005).

DEFINITION 2.21 *A topological space is* weakly regular *if it is semi-regular and, for any non-empty open set u, there exists a non-empty open set v with $v^- \subseteq u$.*

We have

$$X \text{ is normal} \Rightarrow X \text{ is regular} \Rightarrow X \text{ is weakly regular} \Rightarrow X \text{ is semi-regular.}$$

The reverse implications all fail (see Düntsch and Winter, 2005 regarding weak regularity, and Steen and Seebach, 1995 for the other cases).

3.1 Contact

We begin by defining the part-of relation in L_C.

LEMMA 2.22 *Let M be a mereotopology over a weakly regular space X, and let $r_1, r_2 \in M$. Then $r_1 \leq r_2$ if and only if $M \models \phi_{\leq}[r_1, r_2]$, where $\phi_{\leq}(x_1, x_2)$ is the L_C-formula $\forall z(C(x_1, z) \rightarrow C(x_2, z))$.*

Proof If $r_1 \leq r_2$ then $r_1^- \subseteq r_2^-$, so $s^- \cap r_1^- \neq \emptyset$ implies $s^- \cap r_2^- \neq \emptyset$ for any s. Conversely, if $r_1 \not\leq r_2$, by weak regularity, let u be a non-empty, open set such that $u^- \subseteq r_1 \cdot (-r_2)$. Since M is a mereotopology, let $s \in M$ be such that $0 \neq s \subseteq u$. Then $s^- \cap r_1^- \neq \emptyset$, but $s^- \cap r_2^- = \emptyset$. QED

In dealing with mereotopologies over weakly regular spaces, we may therefore write the expression $u \leq v$ in L_C-formulas, as a shorthand for $\phi_{\leq}(u, v)$. It follows that the Boolean constants and functions 0, 1, $+$, $\cdot$ and $-$ are also L_C-definable for mereotopologies over weakly regular spaces, and we again freely employ these symbols in L_C-formulas as a shorthand for their definitions.

We now turn to defining the property of connectedness in L_C. We need some technical lemmas.

LEMMA 2.23 *Let M be a mereotopology over a regular topological space X. If $d \subseteq X$ is closed and $p \notin d$, there exists $r \in M$ such that $p \in r$ and $d \subseteq -r$. In fact, there exist $r, s \in M$ such that $p \in r$, $d \subseteq s$ and $r^- \cap s^- = \emptyset$.*

Proof For the first statement, by T_3-separation, let u, v be disjoint open subsets of X such that $p \in u$ and $d \subseteq v$. Since M is a mereotopology, there exists $r \in M$ such that $p \in r \subseteq u$, whence $d \subseteq v \subseteq X \setminus r^- = -r$. The second

statement follows by two applications of the first: choose $s \in M$ such that $p \in -s$ and $d \subseteq s$; now choose $r \in M$ such that $p \in r$ and $s^- \subseteq -r$. QED

LEMMA 2.24 *Let $r, s \in \mathrm{RO}(X)$ for some topological space X. If $p \in r^-$ and $p \in s$, then $p \in (r \cdot s)^-$.*

Proof Let u be any open set containing p. Then $u \cap s$ is also an open set containing p, whence $(u \cap s) \cap r \neq \emptyset$, since $p \in r^-$. That is, $u \cap (s \cdot r) \neq \emptyset$. QED

LEMMA 2.25 *Let M be a mereotopology over a regular topological space. For all $r_1, r_2 \in M$, $r_1^- \cap r_2^- \cap (r_1 + r_2) \neq \emptyset$ if and only if there exist $r_1', r_2' \in M$ such that $r_1' \leq r_1$, $r_2' \leq r_2$, $r_1'^- \cap r_2'^- \neq \emptyset$ and $(r_1' + r_2')^- \cap (-(r_1 + r_2))^- = \emptyset$.*

Proof The if-direction is immediate. For the only-if-direction, suppose $p \in r_1^- \cap r_2^- \cap (r_1 + r_2)$. By Lemma 2.23, let $s \in M$ be such that $p \in s$ and $(-(r_1 + r_2))^- \subseteq -s$; and let $r_1' = r_1 \cdot s$ and $r_2' = r_2 \cdot s$. By Lemma 2.24, $p \in r_1'^- \cap r_2'^-$, whence r_1' and r_2' have the required properties. QED

LEMMA 2.26 *Let M be a mereotopology which respects components. Then $r \in M$ is connected if and only if $r_1^- \cap r_2^- \cap r \neq \emptyset$ for all non-empty, disjoint $r_1, r_2 \in M$ such that $r_1 + r_2 = r$.*

Proof Suppose r_1 and r_2 are non-empty, disjoint elements of M such that $r_1 + r_2 = r$ and $r_1^- \cap r_2^- \cap r = \emptyset$. By Lemma 2.4 (ii), $r = r_1 \cup r_2$, so that r is not connected. Conversely, suppose r is not connected. Let r_1 be a component of r and let $r_2 = r \setminus r_1$. Since M respects components, $r_1 \in M$. Since $r_1 \subseteq r_1 \cup (r_1^- \cap r_2) \subseteq r_1^-$, $r_1 \cup (r_1^- \cap r_2)$ is connected, whence $r_1^- \cap r_2 = \emptyset$ by maximality of components. Thus, $r_2 = r \setminus r_1^- = r \cdot (-r_1)$. Moreover, since r_1 is open and $r_1 \cap r_2 = \emptyset$, we have $r_1 \cap r_2^- = \emptyset$. Therefore $\emptyset = r_1^- \cap r_2^- \cap (r_1 \cup r_2) = r_1^- \cap r_2^- \cap r$ as required. QED

LEMMA 2.27 *Let M be a mereotopology over a regular topological space X such that M respects components, and let $r \in M$. Then r is connected if and only if $M \models \phi_c[r]$, where $\phi_c(x)$ is the L_C-formula*

$$\forall x_1 \forall x_2 (x_1 > 0 \wedge x_2 > 0 \wedge x_1 \cdot x_2 = 0 \wedge x_1 + x_2 = x \rightarrow \\ \exists x_1' \exists x_2' (x_1' \leq x_1 \wedge x_2' \leq x_2 \wedge C(x_1', x_2') \wedge \neg C(x_1' + x_2', -x))).$$

Proof Lemmas 2.25 and 2.26. QED

Together, Lemmas 2.22 and 2.27 guarantee that, for all mereotopologies over regular topological spaces which respect components, the language L_C is at least as expressive as $L_{c,\leq}$. We take it that all mereotopologies of interest fulfil these conditions: that is, the above reconstructions of the part-whole relation and the property of connectedness in L_C are very robust.

We present a further—and more surprising—demonstration of the expressive power of L_C in mereotopologies defined over $\mathbb{R}^2$. We require the following fact about the topology of Euclidean spaces (Newman, 1964, p. 112, c.f. Proposition 2.19).

PROPOSITION 2.28 *Let d_1 and d_2 be closed sets in $\mathbb{R}^2$, at least one of which is bounded. If $\mathbb{R}^2 \setminus d_1$, $\mathbb{R}^2 \setminus d_2$ and $d_1 \cap d_2$ are all connected, then so is $\mathbb{R}^2 \setminus (d_1 \cup d_2)$.*

LEMMA 2.29 *Let $s_1, s_2, t \in \mathrm{RO}(\mathbb{R}^2)$ such that: (i) either s_1 is bounded or s_2 is bounded; (ii) $-(s_1 + t)$, $-(s_2 + t)$ and t are all connected; and (iii) $s_1^- \cap s_2^- = \emptyset$. Then $-(s_1 + s_2 + t)$ is also connected.*

Proof Set $d_i = (s_i + t)^-$ (for $i = 1, 2$). Thus, the complement of d_i is $-(s_i + t)$ (for $i = 1, 2$), and the complement of $d_1 \cup d_2$ is $-(s_1 + s_2 + t)$. Moreover, since t is connected, so is t^-, whence $d_1 \cap d_2 = (s_1 + t)^- \cap (s_2 + t)^- = (s_1^- \cup t^-) \cap (s_2^- \cup t^-) = (s_1^- \cap s_2^-) \cup t^- = t^-$ is connected. The result follows by Proposition 2.28. QED

Let ϕ_{c} be as defined in Lemma 2.27, and let $\phi_{\mathrm{ub}}(y_1, y_2)$ be the L_C-formula

$$\exists z(\phi_{\mathrm{c}}(-(y_1 + z)) \wedge \phi_{\mathrm{c}}(-(y_2 + z)) \wedge \phi_{\mathrm{c}}(z) \wedge \neg\phi_{\mathrm{c}}(-(y_1 + y_2 + z))).$$

LEMMA 2.30 *Let M be a mereotopology over $\mathbb{R}^2$ such that M respects components and every unbounded element in M includes regions s_1, s_2 and t situated as in Fig. 2.7. Then for all $r \in M$, r is bounded if and only if $M \models \phi_{\mathrm{b}^2}[r]$, where $\phi_{\mathrm{b}^2}(x)$ is the L_C-formula:*

$$\forall y_1 \forall y_2 (y_1 \leq x \wedge y_2 \leq x \wedge \phi_{\mathrm{ub}}(y_1, y_2) \rightarrow C(y_1, y_2)).$$

(The superscript 2 in ϕ_{b^2} refers to the fact that this formula works for mereotopologies over $\mathbb{R}^2$, and not, for example $\mathbb{R}^3$.)

Proof If r does not satisfy $\phi_{\mathrm{b}^2}(x)$ then, by Lemma 2.29, r contains two unbounded regions, so is certainly itself unbounded. Conversely, if r is unbounded, let $s_1, s_2, t \in M$ be subsets of r situated as in Fig. 2.7. Thus, $s_1 \leq r$, $s_2 \leq r$ and $s_1^- \cap s_2^- = \emptyset$, but at the same time, s_1, s_2 satisfies $\phi_{\mathrm{ub}}(y_1, y_2)$,

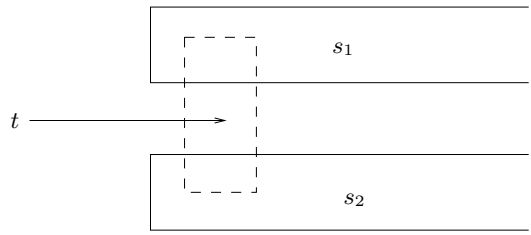


Figure 2.7. Expressing boundedness in L_C: s_1 and s_2 are unbounded to the right.

with t a witness for the existentially quantified z. Hence r does not satisfy $\phi_{\mathrm{b}^2}(x)$. QED

It is simple to verify that the mereotopologies $\mathrm{RO}(\mathbb{R}^2)$, $\mathrm{ROS}(\mathbb{R}^2)$, $\mathrm{ROP}(\mathbb{R}^2)$ and $\mathrm{ROQ}(\mathbb{R}^2)$ satisfy the conditions of Lemma 2.30. Hence, the property of boundedness is L_C-definable in all these mereotopologies. Nevertheless, Lemma 2.30, unlike Lemmas 2.22 and 2.27, has a *fragile* character, in that it depends on a very specific feature of the topological space $\mathbb{R}^2$; in particular, it fails to define boundedness for the corresponding mereotopologies over $\mathbb{R}^3$. We will see in Sec. 6 that boundedness is also L_C-definable in well-behaved mereotopologies over $\mathbb{R}^3$, but we have to go to much more trouble.

3.2 Reconstruction of points

In mereotopologies, the primitive objects—that is, the entities over which variables range—are regions, rather than points; but it is often simple to 'construct' points from regions, and 'simulate' statements about points using statements about regions. One way to construct the point p is as a pair of regions whose closures intersect in the singleton $\{p\}$, as we now proceed to show. (There are also more sophisticated ways, described in Sec. 7.1.)

LEMMA 2.31 *Let M be a mereotopology over a regular topological space, and let $r, s \in M$. Then $r^- \cap s^-$ is a singleton if and only if $M \models \phi_{\bowtie}[r, s]$, where $\phi_{\bowtie}(x_1, x_2)$ is the formula*

$$C(x_1,x_2)\wedge \\ \forall y_1 \forall y_2 (y_1 \leq x_1 \wedge y_2 \leq x_2 \wedge C(y_1, x_2) \wedge C(y_2, x_1) \rightarrow C(y_1, y_2)).$$

Furthermore, if $r^- \cap s^- = \{p\}$ and $t \in M$, then $p \in t$ if and only if $M \models \phi_{\in}[r, s, t]$, where $\phi_{\in}(x_1, x_2, x_3)$ is the formula

$$\exists y_1 (y_1 \leq x_1 \wedge C(y_1, x_2) \wedge \neg C(y_1, -x_3));$$

likewise, $p \in t^-$ *if and only if* $M \models \phi_{\bar{\in}}[r, s, t]$*, where* $\phi_{\bar{\in}}(x_1, x_2, x_3)$ *is the formula*

$$\forall y_1(y_1 \leq x_1 \wedge C(y_1, x_2) \rightarrow C(y_1, x_3)).$$

Proof Routine by Lemmas 2.23 and 2.24. QED

If M is a mereotopology over a topological space X, let us say that M is *complete* if every point in X is the singleton intersection of some pair of regions in M. For example, the mereotopologies $\mathrm{ROP}(\mathbb{R}^n)$, $\mathrm{ROS}(\mathbb{R}^n)$ evidently possess this property; by contrast, $\mathrm{ROQ}(\mathbb{R}^n)$ does not. We might say that, in a complete mereotopology, points can be "simulated" by pairs of regions satisfying the formula $\phi_{\bowtie}$. If M is a complete mereotopology over a regular space, Lemma 2.31 gives us the right to include expressions such as, for example, $x_1 \cap x_2^- \neq \emptyset$ or $\mathcal{F}(x_1) \cap \mathcal{F}(x_2) \subseteq \mathcal{F}(x_3) \cap \mathcal{F}(x_4)$ etc. in L_C-formulas with the obvious interpretation, since such expressions can evidently be replaced by *bona fide* L_C-formulas with the appropriate extension over M.

The following lemma illustrates how easily we can express various topological relations in L_C:

LEMMA 2.32 *Let* $r, s \in \mathrm{ROP}(\mathbb{R}^n)$*. Then* $r^- \cap s^-$ *is connected if and only if* $\mathrm{ROP}(\mathbb{R}^n) \models \phi_{\mathrm{ci}}[r, s]$*, where* $\phi_{\mathrm{ci}}(x, y)$ *is the formula*

$$\forall z \neg(x^- \cap y^- \cap z \neq \emptyset \wedge x^- \cap y^- \cap -z \neq \emptyset \wedge x^- \cap y^- \subseteq z \cup -z).$$

Proof The only-if direction is immediate. So suppose $r^- \cap s^-$ is not connected; we must find a witness for z to show that $\mathrm{ROP}(\mathbb{R}^n) \models \neg\phi_{\mathrm{ci}}[r, s]$. But, by construction of $\mathrm{ROP}(\mathbb{R}^n)$, both r^- and s^- are expressible as finite unions of closed, convex semi-linear sets; and so, therefore, is $r^- \cap s^-$. Since this latter set is not connected, it can be written as $d \cup e$, such that $d \cap e = \emptyset$ and d and e are both finite unions of non-empty, closed, convex semi-linear sets—say, $d = d_1 \cup \cdots \cup d_l$, $e = e_1 \cup \cdots \cup e_m$. Given that such pairs of sets in $\mathbb{R}^n$ can be separated by a hyperplane, we have half-spaces $h_{i,j}$ such that $d_i \subseteq h_{i,j}$ and $e_j \subseteq -h_{i,j}$ for all i, j ($1 \leq i \leq l$, $1 \leq j \leq m$). Then the required witness is

$$t = \sum_{1 \leq i \leq l} \prod_{1 \leq j \leq m} h_{i,j}.$$

QED

3.3 Compactifications

Before discussing the expressive power of $L_{c,\leq}$, we introduce some additional technical material that will be useful throughout this chapter. Recall that a topological space is said to be *locally compact* if every point has a compact neighbourhood. This property "transfers", for Hausdorff spaces, to mereotopologies defined over them:

LEMMA 2.33 *Let M be a mereotopology over a locally compact, Hausdorff space X, and let $p \in X$. Then p is contained within some $r \in M$ such that r^- is compact.*

Proof Let $p \in X$. Assuming X is locally compact, let $d \subseteq X$ be compact and $o \subseteq d$ be open such that $p \in o$. Now let $r \in M$ such that $p \in r \subseteq o \subseteq d$. But a closed subset of a compact set is always compact, and, in a Hausdorff space, every compact set is closed. Therefore $r^- \subseteq d^- = d$ is compact, as required. QED

Let X be a topological space, and let τ denote the collection of open sets of X. Now set $\dot{X} = X \cup \{\infty\}$, where ∞ is some object not in X. For $o \in \tau$, denote by $\dot{o}$ the set

$$\dot{o} = \begin{cases} o \cup \{\infty\} \text{ if } X \setminus o \text{ is compact;} \\ o \text{ otherwise,} \end{cases}$$

and denote by $\dot{\tau}$ the set $\tau \cup \{\dot{o} \mid o \in \tau\}$. Then we can take $\dot{X}$ to be a topological space whose collection of open sets is $\dot{\tau}$. Under this topology (which we always assume), we call $\dot{X}$ the *one-point* (or *Alexandroff*) *compactification* of X. The object ∞ is called the *point at infinity*. The space $\dot{X}$ is always compact. If X is locally compact and Hausdorff, then $\dot{X}$ is also Hausdorff, and hence normal.

NOTATION 2.34 *In this chapter, we denote spheres, open balls and closed balls in Euclidean spaces as follows*

$$\begin{aligned} \mathbf{S}^n &= \{(a_1, \ldots, a_{n+1}) \in \mathbb{R}^{n+1} \mid a_1^2 + \cdots + a_{n+1}^2 = 1\} \\ \mathbf{B}^n &= \{(a_1, \ldots, a_n) \in \mathbb{R}^n \mid a_1^2 + \cdots + a_n^2 < 1\} \\ \mathbf{D}^n &= \{(a_1, \ldots, a_n) \in \mathbb{R}^n \mid a_1^2 + \cdots + a_n^2 \leq 1\}; \end{aligned}$$

and we assume the usual topology on these sets.

(Recall that, by a *d-cell*, we mean any set homeomorphic to the open d-dimensional ball $\mathbf{B}^d$.) In the special cases $X = \mathbb{R}^n$, it is well-known that $\dot{X}$ is homeomorphic to $\mathbf{S}^n$ via the mapping:

$$\begin{aligned} \infty &\mapsto (0, \ldots, 0, 1) \\ (a_1, \ldots, a_n) &\mapsto (a'_1, \ldots, a'_{n+1}), \end{aligned}$$

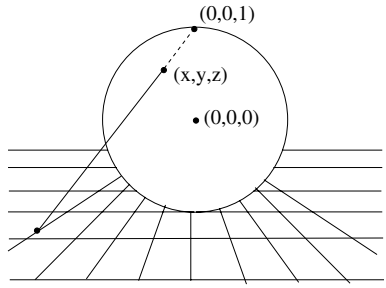


Figure 2.8. Stereographic projection of $\boldsymbol{S}^2$ onto the 1-point compactification of $\mathbb{R}^2$.

where

$$a'_i = 4a_i/(a_1^2 + \cdots + a_n^2 + 4) \qquad \text{for } 1 \leq i \leq n$$
$$a'_{n+1} = (a_1^2 + \cdots + a_n^2 - 4)/(a_1^2 + \cdots + a_n^2 + 4).$$

This mapping may be regarded as a stereographic projection by embedding $\mathbb{R}^n$ in the hyperplane of $\mathbb{R}^{n+1}$ defined in Cartesian geometry by the equation $x_{n+1} = -1$. This projection is depicted for the case $n = 2$ in Fig. 2.8. By way of allusion to this homeomorphism:

NOTATION 2.35 *Let $\mathbb{S}^n$ denote the 1-point compactification of $\mathbb{R}^n$.*

LEMMA 2.36 *Let X be a non-compact topological space. Then the mapping $r \mapsto \dot{r}$ is a Boolean algebra isomorphism from* $\mathrm{RO}(X)$ *to* $\mathrm{RO}(\dot{X})$.

Proof The function $o \mapsto \dot{o}$ is monotone, because a closed subset of a compact set is compact. Let o_1 and o_2 be open subsets of X, with $o = o_1 \cap o_2$. Since $X \setminus o = (X \setminus o_1) \cup (X \setminus o_2)$ is compact if and only if both $(X \setminus o_1)$ and $(X \setminus o_2)$ are compact, we have $\infty \in \dot{o}$ if and only if $\infty \in \dot{o_1} \cap \dot{o_2}$, whence $\dot{o} = \dot{o_1} \cap \dot{o_2}$.

If u is open in X, let u^* denote the pseudo-complement of u in X, and let $(\dot{u})^\star$ denote the pseudo-complement of $\dot{u}$ in $\dot{X}$. We claim that, for any open set u of X, with $v = u^*$, $(\dot{u})^\star = \dot{v}$. By definition, $\dot{v}$ is open in $\dot{X}$, and we have just shown that $\dot{u} \cap \dot{v} = \dot{\emptyset} = \emptyset$. Moreover, if w is any open set in $\dot{X}$ disjoint from $\dot{u}$, then for some open subset w' of X, we have either $w = w'$ or $w = \dot{w'}$. Either way, $u \cap w' = \emptyset$, whence $w' \subseteq v$ and $\dot{w'} \subseteq \dot{v}$ by monotonicity. Hence $\dot{v}$ is the largest open subset of $\dot{X}$ disjoint from $\dot{u}$, i.e. $(\dot{u})^\star = \dot{v}$.

Note that if $r \in \mathrm{RO}(X)$, we have $r = r^{**}$ and $-r = r^*$. Hence $\dot{r}^{\star\star} = \dot{r}$, so that $\dot{r} \in \mathrm{RO}(\dot{X})$. Conversely, if $u' \in \mathrm{RO}(\dot{X})$, then $u' = \dot{u}$, for some open $u \subseteq X$. But then, we have $\dot{u} = \dot{u}^{\star\star} = \dot{x}$, where $x = u^{**}$. Since the function $o \mapsto \dot{o}$ is injective, $u = u^{**}$. That is, $u \in \mathrm{RO}(X)$. QED

LEMMA 2.37 *Let X be a topological space and $o \subseteq X$ open. If o is connected in X, then $\dot{o}$ is connected in $\dot{X}$. Conversely, suppose X is non-compact, and for any closed subsets d_1 and d_2 of X with $X = d_1 \cup d_2$ and $d_1 \cap d_2$ compact, either d_1 is compact or d_2 is compact. If $\dot{o}$ is connected in $\dot{X}$, then o is connected in X.*

Proof Suppose o is open in X. If $\dot{o}$ is not connected in $\dot{X}$, let o_1, o_2 be non-empty open subsets of X such that $\dot{o} = \dot{o}_1 \cup \dot{o}_2$ and $\dot{o}_1 \cap \dot{o}_2 = \emptyset$. Then $o = o_1 \cup o_2$ and $o_1 \cap o_2 = \emptyset$, so o is not connected in X. Conversely, suppose o is not connected in X, so let o_1, o_2 be non-empty open subsets of X such that $o = o_1 \cup o_2$ and $o_1 \cap o_2 = \emptyset$. If $X \setminus o$ is not compact, then neither $X \setminus o_1$ nor $X \setminus o_2$ is compact, so that $\dot{o} = o = o_1 \cup o_2 = \dot{o}_1 \cup \dot{o}_2$ and $\dot{o}_1 \cap \dot{o}_2 = \emptyset$, whence $\dot{o}$ is not connected. If, on the other hand, $X \setminus o$ is compact, by the condition of the lemma, either $X \setminus o_1$ or $X \setminus o_2$ is compact, whence $\dot{o} = o \cup \{\infty\} = o_1 \cup o_2 \cup \{\infty\} = \dot{o}_1 \cup \dot{o}_2$. Moreover, by repeating the first paragraph of the proof of Lemma 2.36, we have $\dot{o}_1 \cap \dot{o}_2 = \dot{\emptyset} = \emptyset$. It follows that $\dot{o}$ is not connected. QED

The well-known Heine-Borel theorem states that, in $\mathbb{R}^n$, a set is compact if and only if it is closed and bounded. It is therefore easy to see that $\mathbb{R}^n$ satisfies the condition of Lemma 2.37.

LEMMA 2.38 *Let $n > 0$ and let M be any mereotopology over $\mathbb{R}^n$. Then the mapping $r \mapsto \dot{r}$ defines a structure isomorphism from M to $\dot{M}$ for the signature $(c, \leq)$: that is, $M \simeq_{c,\leq} \dot{M}$.*

Proof Lemmas 2.36 and 2.37. QED

LEMMA 2.39 *Let X be a locally compact, non-compact topological space and M a mereotopology over X. Define $\dot{M} = \{\dot{r} \mid r \in M\}$. Then $\dot{M}$ is a mereotopology over $\dot{X}$. We call $\dot{M}$ the 1-point compactification of M. If M is finitely decomposable, then so is $\dot{M}$.*

Proof Suppose that $\infty \in \dot{o}$ with o open in X; we show that there exists some $r \in \dot{M}$ such that $\infty \in r \subseteq \dot{o}$. Since M is a mereotopology over X and X is locally compact, Lemma 2.33 gives us a cover of $X \setminus o$ by elements of M whose closures are compact. Since $\infty \in \dot{o}$, $X \setminus o$ is compact, so that this cover has a finite sub-cover, say $r_1, \ldots, r_n$. Let $r = -(r_1 + \cdots + r_n)$. Thus, $X \setminus r = r_1^- \cup \cdots \cup r_n^-$ is compact and includes o, whence r has the required properties. The rest of the Lemma follows from Lemma 2.37. QED

Suppose now that $X = \mathbb{R}^n$ for some $n > 0$, and let M be a mereotopology over $\mathbb{R}^n$ respecting components. Then X satisfies the condition of Lemma 2.37, so by Lemma 2.39, $\dot{M}$ is a mereotopology over $\mathbb{R}^n$ respecting components.

Since $\mathbb{S}^n$ denotes the 1-point-compactification of $\mathbb{R}^n$, the 1-point compactification of $\mathrm{RO}(\mathbb{R}^n)$ is thus $\mathrm{RO}(\mathbb{S}^n)$.

NOTATION 2.40 *Let* $\mathrm{ROS}(\mathbb{S}^n)$ *denote the 1-point compactification of* $\mathrm{ROS}(\mathbb{R}^n)$*, and similarly for* $\mathrm{ROP}(\mathbb{S}^n)$*,* $\mathrm{ROQ}(\mathbb{S}^n)$*.*

It is often more convenient to work with $\mathbb{S}^2$ and $\mathbb{S}^3$ rather than $\mathbb{R}^2$ and $\mathbb{R}^3$. When we need to make the distinction explicit, we refer to elements of $\mathrm{ROP}(\mathbb{R}^n)$ as polytopes (polyhedra, polygons) *in open space* and those of $\mathrm{ROP}(\mathbb{S}^n)$ as polytopes (polyhedra, polygons) *in closed space*. Note that, by Lemma 2.38, the mereotopologies $\mathrm{RO}(\mathbb{R}^n)$, $\mathrm{ROP}(\mathbb{R}^n)$, $\mathrm{ROQ}(\mathbb{R}^n)$ and $\mathrm{ROS}(\mathbb{R}^n)$ certainly all have the same $L_{c,\leq}$-theories as their respective 1-point compactifications.

3.4 Connectedness: the closed plane

We have seen that, over most mereotopologies of interest, the language L_C is as expressive as the language $L_{c,\leq}$. The question therefore arises as to whether a converse reduction is possible. In this section, we show that, for well-behaved mereotopologies over $\mathbb{S}^2$, the answer is positive.

We assume familiarity with basic geometric topology in the plane: for details, see Newman, 1964. Recall in this context that a *Jordan arc* in a topological space X is a homeomorphism from the unit interval $[0, 1]$ into X, and a *Jordan curve* in X, a homeomorphism from the unit circle $\mathbf{S}^1$ into X. The Jordan curve theorem states that the locus of a Jordan curve in $\mathbb{R}^2$ separates $\mathbb{R}^2$ into two residual domains, exactly one of which is bounded. If we regard $\mathbf{S}^1$ as the intersection of the plane $x_1 = 0$ with $\mathbf{S}^2$, the Schönflies Theorem states that a Jordan curve $\gamma : \mathbf{S}^1 \to \mathbb{S}^2$ may be extended to a homeomorphism $\mathbf{S}^2 \leftrightarrow \mathbb{S}^2$. Thus, if γ is a Jordan curve in $\mathbb{S}^2$, the residual domains of $|\gamma|$ are 2-cells in $\mathbb{S}^2$; and if γ is a Jordan curve in $\mathbb{R}^2$, the bounded residual domain of γ is a 2-cell in $\mathbb{R}^2$.

The following concepts are important in understanding the good behaviour of the mereotopologies $\mathrm{ROS}(\mathbb{R}^2)$, $\mathrm{ROP}(\mathbb{R}^2)$ and $\mathrm{ROQ}(\mathbb{R}^2)$.

DEFINITION 2.41 *Let X be a topological space, $u \subseteq X$ and $p, q \in \mathcal{F}(u)$. An* end-cut to p in u *is a Jordan arc in X such that $f(1) = p$ and $f([0, 1[) \subseteq u$. Likewise, a* cross-cut from p to q in u *is a Jordan arc in X such that $f(0) = p$, $f(1) = q$ and $f(]0, 1[) \subseteq u$. Let M be a mereotopology over X. We say that M* has curve-selection *if, for all $r \in M$ and all $p \in \mathcal{F}(r)$, there exists an end-cut in r to p.*

The existence of end-cuts is by no means a universal property of regular open sets in $\mathbb{R}^n$ (for $n > 1$). However, the regions in $\mathrm{ROS}(\mathbb{R}^2)$, $\mathrm{ROP}(\mathbb{R}^2)$ and $\mathrm{ROQ}(\mathbb{R}^2)$ are well-behaved in this regard, as the following results show.

LEMMA 2.42 *Let $r \in \mathrm{ROP}(\mathbb{R}^n)$ and $p \in r^-$. Then there exists a linear function $f : [0,1] \to \mathbb{R}^n$ such that $f(1) = p$ and $f([0,1[) \subseteq r$. If p has rational coordinates, we may choose f so that it has parameters from $\mathbb{Q}$.*

Proof The proposition holds for basic polytopes because their closures are convex. It holds for all polytopes because if $r = r_1 + \cdots r_n$, $r^- = r_1^- \cup \cdots \cup r_n^-$ by Lemma 2.4 (iii). QED

The semi-algebraic case is much more involved. However, we have the following theorem (van den Dries, 1998, Ch. 6, Corollary 1.5; Bochnak et al., 1998 Theorem 2.5.5).

PROPOSITION 2.43 (CURVE-SELECTION LEMMA) *Let S be a semi-algebraic subset of $\mathbb{R}^n$ and $p \in S^-$. Then there exists a continuous semi-algebraic function $f : [0,1] \to \mathbb{R}^n$ such that $f(1) = p$ and $f([0,1[) \subseteq S$.*

Thus, the mereotopologies $\mathrm{ROS}(\mathbb{R}^2)$, $\mathrm{ROP}(\mathbb{R}^2)$ and $\mathrm{ROQ}(\mathbb{R}^2)$ all certainly have curve-selection. Moreover, by making only minor modifications to the relevant arguments, it can be shown that $\mathrm{ROS}(\mathbb{S}^2)$, $\mathrm{ROP}(\mathbb{S}^2)$ and $\mathrm{ROQ}(\mathbb{S}^2)$ all have curve-selection too.

With these preliminaries behind us, we can turn to the expressive power of $L_{c,\leq}$. We note in passing that, since $\leq$ is a primitive of $L_{c,\leq}$, we may write the Boolean operators and constants $+$, $\cdot$, $-$, 0 and 1 in $L_{c,\leq}$-formulas, assuming them to be replaced by their usual definitions. In mereotopologies over the closed plane having curve-selection, we can express the property of being a 2-cell using an $L_{c,\leq}$-formula. To see this, we recall that the Jordan Curve Theorem has the following converse (see Newman, 1964 Ch. VI, Theorem 16.1).

PROPOSITION 2.44 (CONVERSE OF JORDAN'S THEOREM) *Let d be a closed subset of $\mathbb{S}^2$ such that $\mathbb{S}^2 \setminus d$ has two components, and suppose that, for each $p \in d$, and each component o of $\mathbb{S}^2 \setminus d$, there is an end-cut to p in o. Then d is the locus of a Jordan curve.*

Then we have:

LEMMA 2.45 *Let M be any mereotopology over $\mathbb{S}^2$ having curve-selection. Then, for all $r \in M$, r is a 2-cell if and only if r is non-zero and connected with non-zero connected complement—that is, if and only if $M \models \psi_{\mathrm{J}}[r]$, where $\psi_{\mathrm{J}}(x)$ is the $L_{c,\leq}$-formula*

$$c(x) \wedge x > 0 \wedge c(-x) \wedge -x > 0.$$

Proof If $M \models \psi_{\mathrm{J}}[r]$, then $d = \mathcal{F}(r)$ satisfies the conditions of Proposition 2.44, since M has curve-selection. The other direction is immediate. QED

Furthermore:

LEMMA 2.46 *Let M be a mereotopology over $\mathbb{R}^2$ having curve-selection and also satisfying the conditions of Lemma 2.30. Then $r \in M$ is a 2-cell if and only if r satisfies the L_C-formula*

$$\phi_{\mathrm{c}}(x) \wedge x > 0 \wedge \phi_{\mathrm{c}}(-x) \wedge -x > 0 \wedge \phi_{\mathrm{b}^2}(x),$$

where $\phi_{\mathrm{c}}(x)$ and $\phi_{\mathrm{b}^2}(x)$ are as defined in Lemmas 2.27 and 2.30, respectively.

Proof If r satisfies the formula, then the bounded set $\mathcal{F}(r)$ is the locus of a Jordan curve in $\mathbb{S}^2$ and hence in $\mathbb{R}^2$ by the same reasoning as for Lemma 2.45, and since r is the bounded residual domain of this set, it is a 2-cell. The other direction is again immediate. QED

We now proceed to a direct comparison between $L_{c,\leq}$ and L_C. Proposition 2.28 has a closed-plane variant, in which the condition that one of d_1 and d_2 is bounded may be dropped.

PROPOSITION 2.47 *Let d_1 and d_2 be closed sets in $\mathbb{S}^2$. If $\mathbb{S}^2 \setminus d_1$, $\mathbb{S}^2 \setminus d_2$ and $d_1 \cap d_2$ are all connected, then so is $\mathbb{S}^2 \setminus (d_1 \cup d_2)$.*

This leads to a closed-plane variant of Lemma 2.29:

LEMMA 2.48 *Let $s_1, s_2, t \in \mathrm{RO}(\mathbb{S}^2)$ such that: (i) $-(s_1+t)$, $-(s_2+t)$ and t are all connected; and (ii) $s_1^- \cap s_2^- = \emptyset$. Then $-(s_1+s_2+t)$ is also connected.*

Proof As for Lemma 2.29, using Proposition 2.47 in place of Proposition 2.28. QED

We can now state the lemma which ensures that, for certain mereotopologies over the closed plane, $L_{c,\leq}$ is as expressive as L_C.

LEMMA 2.49 *Let M be any finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection, let $\psi_{\mathrm{ub}}(y_1, y_2)$ be the $L_{c,\leq}$-formula*

$$\exists z(c(-(y_1 + z)) \wedge c(-(y_2 + z)) \wedge c(z) \wedge \neg c(-(y_1 + y_2 + z))),$$

and let $\psi_C(x_1, x_2)$ be the $L_{c,\leq}$-formula

$$\exists y_1 \exists y_2 (y_1 \leq x_1 \wedge y_2 \leq x_2 \wedge \psi_{\mathrm{ub}}(y_1, y_2)).$$

Then, for all $r_1, r_2 \in M$, $r_1^- \cap r_2^- \neq \emptyset$ if and only if $M \models \psi_C[r_1, r_2]$.

The if-direction is given by Lemma 2.48. The only-if-direction is given by Lemma 2.83, which we present at the end of Sec. 4.3 (at which point we will

be in a better position to give a succinct proof). In the meanwhile, however, we observe that, for the mereotopologies $\mathrm{ROQ}(\mathbb{S}^2)$, $\mathrm{ROP}(\mathbb{S}^2)$ and $\mathrm{ROS}(\mathbb{S}^2)$ at least, the only-if direction of Lemma 2.49 is almost immediate. (The reader may wish to treat it as an exercise.)

Putting together Lemmas 2.22, 2.27 and 2.49, we see that $L_{c,\leq}$ is exactly as expressive as L_C in well-behaved mereotopologies over the closed plane $\mathbb{S}^2$.

As a final example of the expressiveness of the language L_C, we observe that it can distinguish between $\mathbb{R}^2$ and its 1-point compactification.

THEOREM 2.50 *Let M be any of* $\mathrm{ROS}(\mathbb{R}^2)$, $\mathrm{ROP}(\mathbb{R}^2)$ *or* $\mathrm{ROQ}(\mathbb{R}^2)$. *Then* $M \not\equiv_C \dot{M}$.

Proof Recall the L_C-formula $\phi_{\mathrm{b}^2}(x)$ defined in Lemma 2.30, and expressing the property of being bounded over M. Evidently, $M \not\models \forall x \phi_{\mathrm{b}^2}(x)$. But it is an easy consequence of Lemma 2.48 that $\dot{M} \models \forall x \phi_{\mathrm{b}^2}(x)$. QED

Theorem 2.50 stands in sharp contrast to the situation with the signature $\{c, \leq\}$ reported in Lemma 2.38.

4. Expressiveness of first-order languages in plane mereotopologies

In the previous section, we examined the relative expressive power of the languages L_C and $L_{c,\leq}$ for various mereotopologies, in particular those defined over $\mathbb{R}^2$ and $\mathbb{S}^2$. This section characterizes that expressive power in a more 'absolute' way. We employ the following terminology:

DEFINITION 2.51 *Let X be a topological space and let $\bar{u} = u_1, \ldots, u_n$, $\bar{v} = v_1, \ldots, v_n$ be n-tuples of subsets of X. We say that $\bar{u}$ and $\bar{v}$ are* similarly situated (in X), *and write $\bar{u} \sim_X \bar{v}$, if there is a homeomorphism of X onto itself mapping $\bar{u}$ to $\bar{v}$. If X is clear from context, we omit reference to it, and simply write $\bar{u} \sim \bar{v}$. Now let M be a mereotopology over X and Σ a signature of topological primitives. For any L_Σ-formula ϕ with free variables $\bar{x}$, we say that ϕ is* topologically complete (in M over X) *if any pair of tuples of the appropriate arity satisfying $\phi(\bar{x})$ in M are similarly situated in X.*

Readers familiar with basic geometric topology will recognize that the mereotopologies $\mathrm{ROS}(\mathbb{S}^2)$, $\mathrm{ROP}(\mathbb{S}^2)$ and $\mathrm{ROQ}(\mathbb{S}^2)$ are all (finitely) "triangulable" (in the sense of van den Dries, 1998). Moreover, the observations of Sec. 3.2 strongly suggest that triangulations in these mereotopologies can be combinatorially described using first-order formulas with C as their only primitive. And since combinatorially isomorphic triangulations are similarly situated, it should be entirely unsurprising that every tuple in these mereotopologies satisfies a topologically complete L_C-formula (and hence also a topologically complete

$L_{c,\leq}$-formula). That is: every tuple of regions in any of the mereotopologies $\mathrm{ROS}(\mathbb{S}^2)$, $\mathrm{ROP}(\mathbb{S}^2)$ and $\mathrm{ROQ}(\mathbb{S}^2)$ can be completely topologically described by an L_C-formula (or by an $L_{c,\leq}$-formula). Results of this general kind were proved, independently, by Kuijpers et al., 1995, Papadimitriou et al., 1999 and Pratt and Schoop, 2000, by a variety of methods. Our objective here is a systematic and general investigation of this topic, using an approach which will prove useful in Sections 5 and 7.

4.1 Connected partitions

We have seen that, given a collection Σ of topological primitives, any mereotopology can be regarded as a Σ-structure by interpreting the symbols in Σ in the standard way. And the question then naturally arises as to whether we can obtain a converse to this observation. That is: under what conditions is a given Σ-structure isomorphic to some mereotopology—or perhaps, to some mereotopology belonging to a certain class? Since this question will preoccupy us in the sequel, some of the results below will be presented at a higher level of generality than their immediate applications warrant.

Accordingly, throughout Sections 4.1 and 4.2, $\mathfrak{A}$ shall denote an arbitrary structure interpreting the signature $\{0, 1, +, \cdot, -, c\}$, such that the reduct of $\mathfrak{A}$ to the signature $\{0, 1, +, \cdot, -\}$, is a Boolean algebra. To avoid notational clutter, if $a, b \in A$, we write $0, -a, a + b$ etc., rather than the more correct $0^{\mathfrak{A}}$, $-^{\mathfrak{A}}$(a), $+^{\mathfrak{A}}(a, b)$ etc. In addition, abusing terminology slightly, we call an element $a \in A$ *connected* if $\mathfrak{A} \models c[a]$; and we say that $\mathfrak{A}$ is *finitely decomposable* if, for every $a \in A$, there exist connected elements $a_1, \ldots, a_n$ of $\mathfrak{A}$ such that $a = a_1 + \cdots + a_n$. Of course, in case $\mathfrak{A}$ is a mereotopology M, this usage is consistent with that adopted above. As usual in the context of Boolean algebras, we take a *partition* in $\mathfrak{A}$ to be a tuple of non-zero, pairwise disjoint elements summing to 1. If $\bar{a}$ is any tuple from $\mathfrak{A}$ (not necessarily a partition), and $\bar{b}$ a partition in $\mathfrak{A}$, we say that $\bar{a}$ *can be refined to* $\bar{b}$ if every element of $\bar{a}$ can be written as the sum of (zero or more) elements of $\bar{b}$.

DEFINITION 2.52 *A partition $\bar{a} = a_1, \ldots, a_n$ in $\mathfrak{A}$ such that a_i is connected for all i $(1 \leq i \leq n)$ is called a* connected partition.

Let ψ_{con} denote the $L_{c,\leq}$-sentence

$$\forall x \forall y (c(x) \wedge c(y) \wedge x \cdot y \neq 0 \rightarrow c(x + y)).$$

Thus, ψ_{con} "says" that the sum of two overlapping connected regions is connected.

LEMMA 2.53 *Let M be any mereotopology. Then $M \models \psi_{\mathrm{con}}$.*

Proof A restatement of Lemma 2.4 (iv). QED

CLAIM 2.54 *Suppose $\mathfrak{A}$ is finitely decomposable, and $\mathfrak{A} \models \psi_{\mathrm{con}}$. Then every tuple in $\mathfrak{A}$ can be refined to a connected partition.*

Proof Given elements $a_1, \ldots, a_n$, collect all the non-zero products $b_1, \ldots, b_N$ of the form: $\pm a_1 \cdot \cdots \cdot \pm a_n$. For each j ($1 \leq j \leq N$), let $b_{j,1}, \ldots, b_{j,N_j}$ be connected elements of $\mathfrak{A}$ summing to b_j. If, for any two of these elements, say $b_{j,k}$ and $b_{j,l}$, we have $b_{j,k} \cdot b_{j,l} > 0$, then we can replace them by their sum $b_{j,k} + b_{j,l}$ (which is connected, because $M \models \psi_{\mathrm{con}}$). Proceeding in this way, we obtain the desired refinement. QED

Note that, in particular, every tuple in any finitely decomposable mereotopology can be refined to a connected partition.

Let us restrict attention now to finitely decomposable mereotopologies over $\mathbb{S}^2$ having curve-selection.

LEMMA 2.55 *Let M be a mereotopology over $\mathbb{R}^2$ or $\mathbb{S}^2$ having curve-selection. If r_1, r_2 and r_3 are pairwise disjoint, connected elements of M, then there exist at most two points lying on the frontiers of all three regions.*

Proof We suppose that p_1, p_2 and p_3 are distinct points all lying on the frontiers of r_1, r_2 and r_3 and derive a contradiction. Choose points q_1, q_2, q_3 such that $q_i \in r_i$ ($i = 1, 2, 3$). By curve-selection, draw three end-cuts in r_i, say $\gamma_{i,1}, \gamma_{i,2}$ and $\gamma_{i,3}$ from q_i to p_1, p_2 and p_3, respectively. It is easy to see that, within each r_j ($1 \leq j \leq 3$), the $\gamma_{i,j}$ can be chosen so that they intersect only at q_i. But since the r_j are disjoint, each $\gamma_{i,j}$ intersects any other $\gamma_{i',j'}$ only in p_i or q_i. And it is well known that this is impossible (see the right-hand graph in Fig. 2.10). QED

For $n > 2$, let ψ^n_{sum} denote the $L_{c,\leq}$-formula

$$\forall x_1 \ldots \forall x_n \Big(c(x_1 + \cdots + x_n) \wedge \bigwedge_{1 \leq i \leq n} c(x_i) \rightarrow \bigvee_{2 \leq i \leq n} c(x_1 + x_i) \Big).$$

(The formula ψ_{sum} of Example 2.18 is just ψ^3_{sum}.) Thus, ψ^n_{sum} "says" that, if n connected regions have a connected sum, the first must form a connected sum with at least one of the others.

LEMMA 2.56 *Let M be a finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection. Then $M \models \psi^n_{\mathrm{sum}}$ for all $n > 1$.*

Proof Let $r_1, \ldots, r_n$ be connected with $r_1 + \cdots + r_n$ also connected. Assume first that the r_i are pairwise disjoint. Let $p \in r_1$ and $q \in r_2 + \cdots + r_n$. By the connectedness of $r_1 + \cdots + r_n$, draw a Jordan arc γ from p to q lying within $r_1 + \cdots + r_n$. By Lemma 2.55, only finitely many points can lie on the frontiers of more than two of the r_i, and we may certainly ensure that γ avoids all such points. By renumbering if necessary, we may assume that γ visits a

point $p \in r_1^- \cap r_2^- \cap (r_1 + \ldots + r_n)$. But by the construction of γ, $p \notin r_i^-$ for all $i > 2$, whence $p \in -r_i$ for all such i. Therefore, $p \in r_1^- \cap r_2^- \cap (r_1 + r_2)$, whence $r_1 + r_2$ is connected. Finally, we relax the assumption that the r_i are pairwise disjoint. Since M is finitely decomposable, we have that each element of $\bar{r}$ is the sum of zero or more members of a tuple $\bar{s}$ of pairwise disjoint connected elements with the same sum. The result then follows easily by repeated applications of Lemma 2.53. QED

In the sequel, we abbreviate the formula

$$x_1 + x_2 = x \wedge x_1 > 0 \wedge x_2 > 0 \wedge x_1 \cdot x_2 = 0 \wedge c(x_1) \wedge c(x_2)$$

by $x_1 \oplus x_2 = x$; thus, $x_1 \oplus x_2 = x$ "says" that x can be partitioned into non-zero, disjoint connected regions x_1 and x_2. Now let ψ_{break} denote the $L_{c,\leq}$-formula

$$\begin{aligned}\forall x \forall y_1 \forall y_2 \Big(\big(& c(x) \wedge c(y_1) \wedge c(y_2) \wedge c(x + y_1) \wedge \\ & c(x + y_2) \wedge x \cdot y_1 = 0 \wedge x \cdot y_2 = 0 \wedge x \neq 0 \big) \rightarrow \\ \exists x_1 \exists x_2 \big(& x_1 \oplus x_2 = x \wedge c(x_1 + y_1) \wedge c(x_1 + y_2) \wedge c(x_2 + y_1) \wedge c(x_2 + y_2) \big) \Big).\end{aligned}$$

Thus, ψ_{break} "says" that, if r, s_1, s_2 are connected regions such that r is non-zero, disjoint from s_1 and s_2, and forms a connected sum with both s_1 and s_2, then r can be partitioned into connected, non-zero regions r_1, r_2 such that each of r_1 and r_2 forms a connected sum with each of s_1 and s_2. Fig. 2.9a illustrates this configuration; note that $-r$ need not be connected.

LEMMA 2.57 *Let M be a finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection. Then $M \models \psi_{\text{break}}$.*

Proof Let r, s_1, s_2 be as above. We may assume that s_1 and s_2 are non-zero, since otherwise, similar or easier arguments apply. Refer to Fig. 2.9b. For $i = 1, 2$, since $r + s_i$ is connected, by Lemma 2.4 (ii), $r^- \cap s_i^- \cap (r + s_i) \neq \emptyset$. In fact, since the removal of finitely many points from the open set $r + s_i$ does not disconnect it, we can choose four distinct points p_i, q_i ($i = 1, 2$) such that $p_i, q_i \in r^- \cap s_i^- \cap (r + s_i)$. Since M has curve-selection and r is connected it is easy to see that, by exchanging q_1 and q_2 if necessary, we can draw cross-cuts γ from p_1 to p_2 and δ from q_1 to q_2 such that $|\gamma|$ and $|\delta|$ are disjoint. Since $\mathbb{S}^2$ is normal and M a mereotopology, we can cover $|\delta|$ with elements of M whose closures are disjoint from $|\gamma|$. By compactness of $|\delta|$, this cover has a finite subcover, $t_1, \ldots, t_N$, say. Let $t = r \cdot (t_1 + \cdots + t_N)$; evidently, q_1 and q_2 lie on the frontier of the same component t' of t. Likewise, p_1 and p_2 lie on the frontier of the same component of $r \cdot -t'$: call this component $r_1 \in M$, and let $r_2 = r \cdot -r_1$. It is easy to check that r_1 and r_2 have the required properties. QED

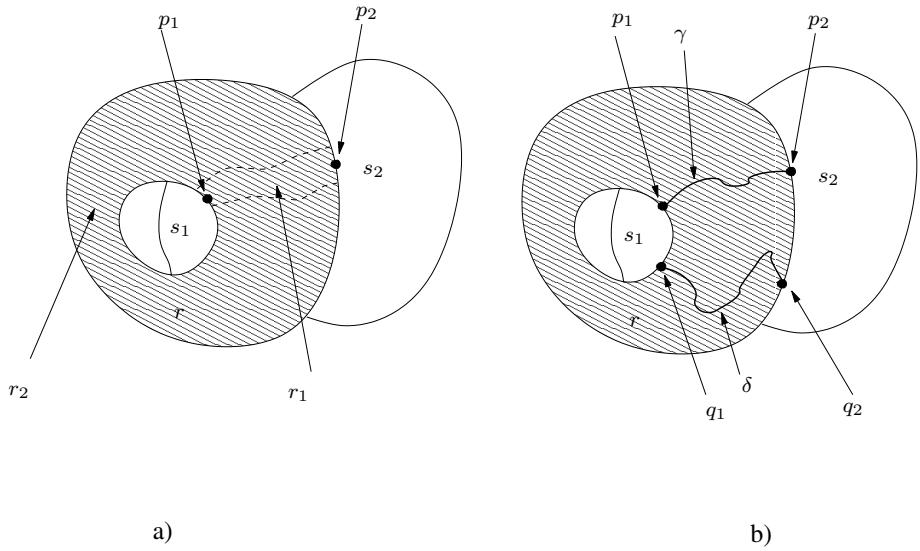


Figure 2.9. The configuration of ψ_{break}.

4.2 Neighbourhood graphs

As before, $\mathfrak{A}$ shall denote an arbitrary structure interpreting the signature $\{0, 1, +, \cdot, -, c\}$, such that the reduct of $\mathfrak{A}$ to $\{0, 1, +, \cdot, -\}$ is a Boolean algebra. Recall the notion of connected partition introduced in Definition 2.52.

DEFINITION 2.58 *Let $\bar{a} = a_1, \ldots, a_n$ be a connected partition in $\mathfrak{A}$. We say that $\bar{a}$ is a* c^h-partition *if, for every $I \subseteq \{1, \ldots, n\}$ such that $|I| < h$, the element $(-\sum_{i \in I} a_i)$ is connected.*

If $\mathfrak{A} \models c(1)$, then a c^1-partition in $\mathfrak{A}$ is the same thing as a connected partition. Furthermore, if $\mathfrak{A}$ is in fact a mereotopology over $\mathbb{S}^2$ having curve-selection, then, by Lemma 2.45, a c^2-partition in $\mathfrak{A}$ is the same thing as a partition consisting entirely of 2-cells. It is c^3-partitions, however, that will mainly preoccupy us in the sequel.

We assume familiarity with basic graph theory: for details, see Diestel, 1991 Ch. 1. Recall in this context that a *graph* is a pair $G = (V, E)$ where V is a finite set (called *vertices*) and E is a set of 2-element subsets of V (called *edges*). We denote V by $V(G)$ and E by $E(G)$. Note that, on this definition, graphs have no "loops" or "multiple edges". If G is a graph and U a proper subset of $V(G)$, we denote by $G \setminus U$ the result of *deleting* all the nodes in U from G; and if $e = (v, v') \in E(G)$, we denote by G/e the result of *contracting* G by merging v and v' into a single (new) node v'', such that (v'', w) is an edge of G/e just in case either (v, w) or (v', w) is an edge of G. If a graph H can be

obtained from G by a sequence of deletions and contractions, then H is said to be a *minor* of G. Finally we take the terms *path*, *cycle*, *connected*, *component* to be defined in the standard way. In particular, recall that, for $h > 0$, G is said to be h-connected if $G \setminus U$ is connected for every $U \subseteq G$ such that $|U| < h$.

DEFINITION 2.59 *Let $\bar{a} = a_1, \ldots, a_n$ be a tuple from $\mathfrak{A}$. If $a_i + a_j$ is connected for $1 \leq i < j \leq n$, we say that a_i and a_j are* neighbours. *The* neighbourhood graph *of $\bar{a}$, denoted $N_{\bar{a}}$, is the graph with nodes $\{a_1, \ldots, a_n\}$ and edges $\{(a_i, a_j) \mid a_i$ and a_j are neighbours$\}$.*

CLAIM 2.60 *Suppose $\mathfrak{A} \models \psi_{\text{con}}$ and $\mathfrak{A} \models \psi^n_{\text{sum}}$ for all $n > 2$. Let $\bar{a} = a_1, \ldots, a_n$ be a tuple of connected elements of $\mathfrak{A}$, such that $a_{n-1} + a_n$ is connected. Let $\bar{a}' = a_1, \ldots, a_{n-2}, (a_{n-1} + a_n)$. Then $N_{\bar{a}'} = N_{\bar{a}}/(a_{n-1}, a_n)$.*

Proof For $1 \leq j < n-1$, $a_j + (a_{n-1} + a_n)$ is connected if and only if $a_j + a_{n-1}$ is connected or $a_j + a_n$ is connected. QED

CLAIM 2.61 *Suppose $\mathfrak{A} \models \psi_{\text{con}}$ and $\mathfrak{A} \models \psi^n_{\text{sum}}$ for all $n > 2$. Let $\bar{a} = a_1, \ldots, a_n$ be a tuple of connected elements of $\mathfrak{A}$, with $a = a_1 + \ldots + a_n$. Then a is connected if and only if $N_{\bar{a}}$ is a connected graph.*

Proof The if-direction follows easily from the fact that $\mathfrak{A} \models \psi_{\text{con}}$. For the only-if direction, note that the claim is trivial if $n = 1$, so assume $n > 1$, and that the claim holds for tuples of fewer than n elements. Since $\mathfrak{A} \models \psi^n_{\text{sum}}$ there exists i ($1 \leq i < n$) such that a_i and a_n are neighbours. By renumbering if necessary, assume $i = n-1$, and let $\bar{a}'$ be as in Claim 2.60, so that $N_{\bar{a}'} = N_{\bar{a}}/(a_{n-1}, a_n)$. But $N_{\bar{a}'}$ is connected by inductive hypothesis, whence $N_{\bar{a}}$ is connected too.

QED

CLAIM 2.62 *Suppose $\mathfrak{A} \models \psi_{\text{con}}$ and $\mathfrak{A} \models \psi^n_{\text{sum}}$ for all $n > 2$. Let $\bar{a}$ be a connected partition in $\mathfrak{A}$, and let $h \geq 1$. Then $\bar{a}$ is a c^h-partition if and only if $N_{\bar{a}}$ is an h-connected graph.*

Proof Immediate by Claim 2.61. QED

CLAIM 2.63 *Suppose $\mathfrak{A} \models c(1)$, $\mathfrak{A} \models \psi_{\text{con}}$, $\mathfrak{A} \models \psi^n_{\text{sum}}$ for all $n > 2$, and $\mathfrak{A} \models \psi_{\text{break}}$. Then every connected partition in $\mathfrak{A}$ can be refined to a c^3-partition.*

Proof We make free use of Claim 2.61. Let $\bar{a}$ be a connected partition. We show first that $\bar{a}$ can be refined to a c^2-partition. Choose an element a of $\bar{a}$ such that the number k of components of the graph $N_{\bar{a}} \setminus \{a\}$ is maximal. And let there be $m > 0$ elements a for which this maximum value is attained. If $\bar{a}$

is not already a c^2-partition, then $k > 1$. Let H_1, H_2 be distinct components of $N_{\bar{a}} \setminus \{a\}$. Since $N_{\bar{a}}$ is connected, there exist $b_1 \in H_1$, $b_2 \in H_2$ such that $a + b_1$ and $a + b_2$ are connected. Since $\mathfrak{A} \models \psi_{\text{break}}$, let a_1, a_2 be non-zero, connected, disjoint elements summing to a with $a_1 + b_1$, $a_1 + b_2$, $a_2 + b_1$ and $a_2 + b_2$ all connected; and let $\bar{b}$ be the connected partition which results from replacing a by a_1 and a_2. Evidently, for $i = 1, 2$, $N_{\bar{b}} \setminus \{a_i\}$ has strictly fewer than k components. That is, the number of elements b in $\bar{b}$ such that $N_{\bar{b}} \setminus \{b\}$ has k components is strictly less than m. Proceeding in this way, we eventually obtain a c^2-partition.

Now let $\bar{a}$ be a c^2-partition. We show that $\bar{a}$ can be refined to a c^3-partition. If $\bar{a}$ is not a c^3-partition, choose a pair of distinct elements a and a' such that the number k of components of the graph $N_{\bar{a}} \setminus \{a, a'\}$ is maximal; and let there be $m > 0$ unordered pairs (a, a') for which this maximum value is attained. Let H_1, H_2 be distinct components of $N_{\bar{a}} \setminus \{a, a'\}$. Since $\bar{a}$ is a c^2-partition, there exist $b_1 \in H_1$, $b_2 \in H_2$ such that $a + b_1$ and $a + b_2$ are connected. And since $\mathfrak{A} \models \psi_{\text{break}}$, let a_1, a_2 be non-zero, connected, disjoint elements summing to a with $a_1 + b_1$, $a_1 + b_2$, $a_2 + b_1$ and $a_2 + b_2$ all connected; and let $\bar{b}$ be the connected partition which results from replacing a by a_1 and a_2. Evidently, for $i = 1, 2$, $N_{\bar{b}} \setminus \{a_i, a'\}$ has strictly fewer than k components. Moreover, suppose a'' is any other element of $\bar{a}$ (distinct from a and a') such that $N_{\bar{a}} \setminus \{a, a''\}$ also has k components. We claim that $N_{\bar{b}} \setminus \{a_1, a''\}$ and $N_{\bar{b}} \setminus \{a_2, a''\}$ cannot both have k components. Working for the moment on this assumption, we see that the number of pairs b, b' in $\bar{b}$ such that $N_{\bar{b}} \setminus \{b, b'\}$ has k components is strictly less than m. Proceeding in this way, we eventually obtain a c^3-partition.

It remains only to verify that the graphs $N_{\bar{b}} \setminus \{a_1, a''\}$ and $N_{\bar{b}} \setminus \{a_2, a''\}$ encountered above do not both have k components. If $a \in A$ and $B \subseteq A$, let us say that a is a neighbour of B if a is a neighbour of some element of B. Let the components of $N_{\bar{a}} \setminus \{a, a''\}$ be $H_1, \ldots H_k$. Since $\bar{b}$ is a c^2-partition, we have that, for all i ($1 \leq i \leq k$), a is a neighbour of H_i, and therefore either a_1 or a_2 is a neighbour of H_i. Hence, we can re-order the H_i if necessary so that, for some p, q with $0 \leq p < q \leq k + 1$, a_1 is a neighbour of H_i if and only if $i < q$ and a_2 is a neighbour of H_i if and only if $p < i$. Thus, the components of $N_{\bar{b}} \setminus \{a_1, a''\}$ are $H_1, \ldots, H_p, (\{a_2\} \cup H_{p+1} \cdots \cup H_k)$, and the components of $N_{\bar{b}} \setminus \{a_2, a''\}$ are $(\{a_1\} \cup H_1 \cdots \cup H_{q-1}), H_q, \ldots, H_k$. If these number k in each case, we have $p = k - 1$ and $q = 2$. But a' lies in one of the H_i, and a_1 and a_2 were chosen so that they are both neighbours of this a'. Hence a_1 and a_2 are both neighbours of H_i, whence $p < q - 1$. This yields $k \leq 1$, contradicting our assumption that $\bar{a}$ is not a c^3-partition. QED

We finish with a technical result which will be required later.

DEFINITION 2.64 *If $\bar{a} = a_1, \ldots, a_N$ is a connected partition in $\mathfrak{A}$ such that, for any neighbour a_j of a_i, $-(a_i + a_j)$ is connected, we say that $\bar{a}$ is* radial about a_i.

Note incidentally that a c^3-partition is radial about each of its members.

CLAIM 2.65 *Suppose $\mathfrak{A} \models c(1)$, $\mathfrak{A} \models \psi_{\text{con}}$, $\mathfrak{A} \models \psi^n_{\text{sum}}$ for all $n > 2$, and $\mathfrak{A} \models \psi_{\text{break}}$. Let $n > 1$ and let $\bar{a} = a_1, \ldots, a_n$ be a connected partition in $\mathfrak{A}$ with $-a_1$ connected. Then $\bar{a}$ can be refined to a c^2-partition $a_1, b_2, \ldots, b_N$, radial about a_1, in which a_1 has at least three neighbours.*

Proof Similar to the above. QED

We conclude with a further corollary of Claim 2.61. We employ the following fact from graph theory, whose proof we leave to the reader.

PROPOSITION 2.66 *If G is a finite 2-connected graph of order at least 2, and $v \in V(G)$, then there exists a $w \in V(G)$ such that $\{v, w\} \in E(G)$, and the removal of both v and w from G leaves a connected graph.*

COROLLARY 2.67 *Let M be a finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection, and let $\bar{r} = r_1, \ldots, r_n$ be a partition in M consisting entirely of 2-cells. Then, by re-numbering if necessary, we have, for all k $(1 \leq k < n)$, $r_1 + \cdots + r_k$ is a 2-cell.*

That is: partitions of the closed plane into 2-balls are always 'shellable'. The analogous result for three-dimensional space fails (Rudin, 1958).

4.3 Partition graphs

We now prove that, if $\bar{r}$ is a c^3-partition in a finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection, then the neighbourhood graph of $\bar{r}$ fixes its topological properties completely.

We assume familiarity with the basic theory of plane graphs: for details, see Diestel, 1991 Ch. 4. In this context, suppose that $e \subseteq \mathbb{S}^2$ is the locus of a Jordan arc. Then e has two *endpoints*; all other points are called *interior points*, and we denote the set of these interior points by (e). (Of course, e is not the *topological* interior of the set e in $\mathbb{S}^2$; but no confusion should arise in this regard.) A *plane graph* is a pair $G = (V, E)$, where V is a finite subset of $\mathbb{S}^2$ and E is a collection of sets $e \subseteq \mathbb{S}^2$ such that e is the locus of a Jordan arc, satisfying the following conditions for all $v \in V$ and all $e, e' \in E$:

1 if $e \in E$ and p is an endpoint of e, then $p \in V$;

2 $v \notin (e)$, and if $e \neq e'$ then $(e) \cap (e') = \emptyset$;

3 if $e \neq e'$, then e and e' do not join the same pair of endpoints.

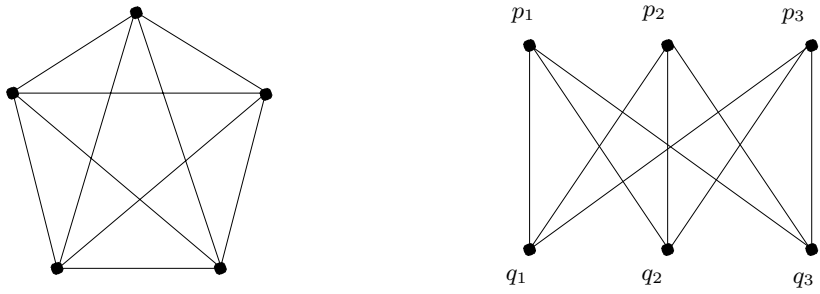


Figure 2.10. The non-planar graphs K^5 and $K_{3,3}$.

The elements of V are called the *vertices* of G, and the elements of E, the *edges* of G; an edge $e \in E$ is said to *join* the vertices at its endpoints. We denote V by $V(G)$, E by $E(G)$ and $V \cup \bigcup E$ by $|G|$. The components of $\mathbb{S}^2 \setminus |G|$ are called the faces of G, and we denote the set of these faces by $F(G)$. A plane graph is *semi-algebraic* if its edges are the loci of semi-algebraic Jordan arcs; similarly for the terms *piecewise linear* and *rational piecewise linear*. Notice that, on our definition, plane graphs have no "loops" or "multiple edges". (Some authors prefer the term *simple graph*.) A plane graph will be regarded as an abstract graph in the obvious way, and we carry over notation and terminology accordingly. Conversely, if $G = (V, E)$ is an abstract graph, a *drawing* of G is a plane graph $G' = (V', E')$ for which there exists a function ϵ mapping V 1–1 onto V' and E 1–1 onto E' such that for all $(v, v') \in E$, $\epsilon((v, v'))$ joins $\epsilon(v)$ and $\epsilon(v')$. We call ϵ an *embedding*. If G has a drawing, G is *planar*. Not all abstract graphs are planar, of course: the graphs K^5 and $K_{3,3}$ illustrated in Fig. 2.10 are familiar non-planar graphs. Indeed, this fact has a converse:

PROPOSITION 2.68 (KURATOWSKI, WAGNER) *A graph is planar if and only if it has no minor isomorphic to either K^5 or $K_{3,3}$.*

We further assume familiarity with the notion of *duality* for plane graphs. Let G and G' be plane graphs. We say that G' is a *geometrical dual* of G if there are bijections $f_F : F(G) \to V(G')$ and $f_E : E(G) \to E(G')$ such that, for all $f \in F(G)$ and $e \in E(G)$:

1 $f_F(f) \in f$;

2 $f_E(e) \cap e$ is a single point interior to both $f_E(e)$ and e, and $f_E(e) \cap e' = \emptyset$ for all $e' \neq e$.

In our terminology, not every plane graph has a dual, because we do not allow graphs to contain loops or multiple edges. However, we rely below on the following sufficient condition (Wilson, 1979, p. 76).

PROPOSITION 2.69 *Every 3-connected plane graph has a dual.*

The following fact is also well-known.

LEMMA 2.70 *Let G and G' be connected plane graphs such that G' is a geometrical dual of G. Then there is a bijection $f_V : V(G) \to F(G')$ such that, for all $v \in V(G)$, $v \in f_V(v)$. Hence, G is a dual of G'.*

Proof Every face of G' contains at least one vertex of G by construction; it contains at most one by Euler's formula $|F(G)| - |E(G)| + |V(G)| = 2$ applied to G and G'. QED

Finally, duals are unique, in the following sense (Diestel, 1991, p. 88).

PROPOSITION 2.71 *Let G be a plane graph and let G' and G'' be plane graphs which are both geometrical duals of G. Then there is a homeomorphism $h : \mathbb{S}^2 \to \mathbb{S}^2$ mapping G' to G''. In fact, h can be chosen such that, for all $v \in G$, if f' and f'' are the faces of G' and G'', respectively, containing v, then h maps f' to f''.*

Now let us apply these ideas to the graphs whose faces are c^3-partitions in well-behaved, closed-plane mereotopologies.

LEMMA 2.72 *Let X be a topological space, and let r, s be disjoint elements of* $\mathrm{RO}(X)$ *with $p \in \mathcal{F}(r) \setminus \mathcal{F}(s)$. Then $p \in \mathcal{F}(-(r+s))$.*

Proof By Lemma 2.4 (ii), $p \notin r + s$. QED

LEMMA 2.73 *Let M be a mereotopology over $\mathbb{S}^2$ having curve-selection, and let $\bar{r} = r_1, \ldots, r_n$ be a c^3-partition in M. For all i, j ($1 \leq i < j \leq n$), $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is connected.*

Proof We may assume that $n \geq 3$. Since $\bar{r}$ is certainly a c^2-partition, every $\mathcal{F}(r_i)$ ($1 \leq i \leq n$) is a Jordan curve by Lemma 2.45. Suppose, for contradiction, that $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is not connected, and let $p, q \in \mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ be separated in $\mathcal{F}(r_i)$ by $\{p', q'\} \subseteq \mathcal{F}(r_i) \setminus \mathcal{F}(r_j)$. By Lemma 2.72, $p', q' \in \mathcal{F}(-(r_i + r_j))$, so that, by the connectedness of $-(r_i + r_j)$, we can draw a cross-cut γ' (Definition 2.41) from p' to q' in $-(r_i + r_j) \subseteq -r_i$. By the connectedness of r_j, we can likewise draw a cross-cut γ from p to q in $r_j \subseteq -r_i$. But $-r_i$ is a 2-cell, whence γ and γ' are easily seen to intersect at an interior point, which is impossible, since $r_j \cap -(r_i + r_j)$ is empty. QED

LEMMA 2.74 *Let M be a mereotopology over $\mathbb{S}^2$ having curve-selection, and let $\bar{r} = r_1, \ldots, r_n$ ($n \geq 4$) be a c^3-partition in M. Then there exists a unique plane graph G drawn in $\mathbb{S}^2$ such that the collection of sets $\{r_1, \ldots, r_n\}$ are*

exactly $F(G)$ *and the collection of sets* $\{\mathcal{F}(r_i) \cap \mathcal{F}(r_j) \mid 1 \leq i < j \leq n, r_i + r_j$ *is connected*$\}$ *are exactly* $E(G)$.

Proof Let i, j, k be distinct integers in the range $[1, n]$. Since $\bar{r}$ is a c^3-partition, $r_j^- \cup r_k^- = \mathbb{S}^2 \setminus -(r_j + r_k)$ does not separate the non-empty sets r_i and $-(r_i + r_j + r_k)$, whence $\mathcal{F}(r_i) \cap (\mathcal{F}(r_j) \cup \mathcal{F}(r_k))$ is not the whole of the Jordan curve $\mathcal{F}(r_i)$. And since, by Lemma 2.73, $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is a connected subset of $\mathcal{F}(r_i)$, $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is either a point or the locus of a Jordan arc. Indeed, $\mathcal{F}(r_i)$ must include at least three Jordan arcs of the form $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ for various j distinct from i. Let the vertices of G be the endpoints of all Jordan arcs of the form $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$, and let the edges of G be the segments of the various $\mathcal{F}(r_i)$ connecting them. To show that G is a plane graph, we must establish that if $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is a Jordan arc γ, then for all k ($1 \leq k \leq n$) with $k \neq i, j$, $\mathcal{F}(r_k)$ contains no interior points of γ. For otherwise, let $p' \in \mathcal{F}(r_k)$ be an interior point of γ, and pick any $q' \in \mathcal{F}(r_i) \setminus \mathcal{F}(r_j)$. Then $p' \in \mathcal{F}(-(r_i + r_j))$ and also, by Lemma 2.72, $q' \in \mathcal{F}(-(r_i + r_j))$. If we now choose p and q in $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ separating p' and q' on the Jordan curve $\mathcal{F}(r_i)$, the derivation of a contradiction proceeds as in Lemma 2.73. Hence no point of $\mathcal{F}(r_k)$ is an interior point of γ, as required. Moreover, no two Jordan arcs in E can have the same end-points, since $\bar{r}$ is a c^3-partition. It follows that G is a plane graph as required. Evidently, $F(G) = \{r_1, \ldots, r_n\}$ and $E(G')$ is the collection of sets $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ for $1 \leq i < j \leq n$ which are Jordan arcs.

It therefore remains only to show that $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is a Jordan arc if and only if $r_i + r_j$ is connected. Note that $r_1 \cup r_2$ is trivially not connected. By Lemma 2.4 (ii), $r_i \cup r_j \subseteq r_i + r_j \subseteq r_i \cup r_j \cup (\mathcal{F}(r_i) \cap \mathcal{F}(r_j))$, and the removal of a single point from a connected, open set does not render it disconnected. Hence, if $r_i + r_j$ is connected, $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is neither empty nor a singleton, and hence is a Jordan arc. Conversely, suppose $\mathcal{F}(r_i) \cap \mathcal{F}(r_j)$ is a Jordan arc. We have already shown that, if p is an interior point of this arc, $p \notin \cup_{k \neq i,j} r_k^- = (\sum_{k \neq i,j} r_k)^-$. That is, $p \in -\sum_{k \neq i,j} r_k = r_i + r_j$. Hence $\mathcal{F}(r_i) \cap \mathcal{F}(r_j) \cap (r_i + r_j)$ is non-empty, whence $r_i + r_j$ is connected. QED

DEFINITION 2.75 *Let* M *be a mereotopology over* $\mathbb{S}^2$ *having curve-selection, and let* $\bar{r} = r_1, \ldots, r_n$ ($n \geq 4$) *be a* c^3*-partition in* M. *We call the unique plane graph* G *satisfying the conditions of Lemma 2.74* the partition graph of $\bar{r}$.

Warning: the neighbourhood graph and the partition graph of a c^3-partition are not the same sort of thing. The former is an *abstract* graph whose nodes are regions and whose edges are pairs of regions; the latter is a *plane* graph, whose nodes are points and whose edges are the loci of Jordan arcs.

LEMMA 2.76 *Let M be a mereotopology over $\mathbb{S}^2$ having curve-selection, let $\bar{r} = r_1, \ldots, r_n$ $(n \geq 4)$ be a c^3-partition in M, and let G be its partition graph. Then there is a plane embedding ϵ of $N_{\bar{r}}$ such that $\epsilon(N_{\bar{r}})$ is a geometrical dual of G and, for all i, $(1 \leq i \leq n)$, $\epsilon(r_i) \in r_i$.*

Proof Almost immediate from the definition of partition graph. QED

From Claim 2.62, c^3-partitions have 3-connected neighbourhood graphs. But 3-connected graphs have the crucial property that all their drawings are topologically the same.

PROPOSITION 2.77 (WHITNEY) *Let G and G' be 3-connected plane graphs and $f : G \to G'$ a graph isomorphism. Then f can be extended to a homeomorphism $h : \mathbb{S}^2 \to \mathbb{S}^2$.*

Let M be a finitely decomposable mereotopology over $\mathbb{S}^2$, and let $\bar{r} = r_1, \ldots, r_n$ and $\bar{s} = s_1, \ldots, s_n$, be n-tuples from M. We are interested in the case where the mapping $r_i \mapsto s_i$ is a graph isomorphism from $N_{\bar{r}}$ to $N_{\bar{s}}$—that is, where, for all i, j, $(1 \leq i < j \leq n)$, $r_i + r_j$ is connected if and only if $s_i + s_j$ is connected. We say in this case that $\bar{r}$ and $\bar{s}$ *have the same neighbourhood structure*.

THEOREM 2.78 *Let M be a finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection. Then any two c^3-partitions in M having the same neighbourhood structure are similarly situated in $\mathbb{S}^2$.*

Proof It is straightforward to verify that, if $n \leq 3$, all n-element c^3-partitions in M are similarly situated in $\mathbb{S}^2$. Thus, we may assume that $n \geq 4$. Let $\bar{r} = r_1, \ldots, r_n$ and $\bar{s} = s_1, \ldots, s_n$ be c^3-partitions with the same neighbourhood structure, and let G and H be their respective partition graphs. By Lemma 2.76, let G^* and H^* be embeddings of $N_{\bar{r}}$ and $N_{\bar{s}}$, geometrically dual to G and H, respectively, let p_i be the vertex of G^* contained in r_i and let q_i be the vertex of H^* contained in s_i for all i $(1 \leq n)$. Hence, there is a graph isomorphism $f : G^* \to H^*$ mapping p_i to q_i. Since G^* and H^* are 3-connected, Proposition 2.77 guarantees that f can be extended to a homeomorphism $h : \mathbb{S}^2 \to \mathbb{S}^2$. Then $h(G)$ and H are both geometrical duals of the plane graph $h(G^*) = H^*$, such that, for all i $(1 \leq i \leq n)$ the faces $h(r_i)$ and s_i contain the vertex $h(p_i) = q_i$. By Proposition 2.71, let h' be a homeomorphism mapping $h(G)$ to H such that $h'(h(r_i)) = s_i$. Thus, $\bar{r}$ and $\bar{s}$ are similarly situated. QED

We finish this discussion of partition graphs with some "obvious" lemmas concerning connected partitions in ROP$(\mathbb{S}^2)$ and related mereotopologies. Readers irritated by proofs of such evident truths may skip to Theorem 2.82.

LEMMA 2.79 *Let G be a plane graph such that G has no isolated vertices, and every edge of G lies on the boundary of (at least) 2 faces of G. Then the*

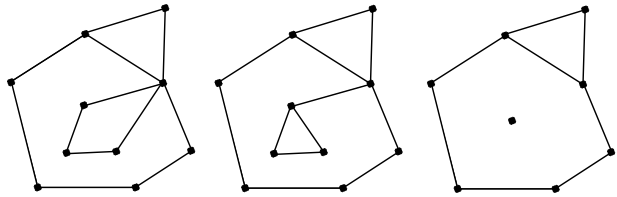


Figure 2.11. Only the left-hand graph defines a connected partition in $\mathrm{RO}(\mathbb{S}^2)$.

members of $F(G)$ are regular open, and form a connected partition in $\mathrm{RO}(\mathbb{S}^2)$. *Moreover, if G' is another such plane graph, with $|G| \subseteq |G'|$, then, for every $f \in F(G)$, $f = \sum\{f' \in F(G') \mid f' \subseteq f\}$.*

Proof Let $G = (V, E)$, and suppose $f \in F(G)$ and $p \in \mathcal{F}(f)$. Since G has no isolated vertices, there exists $e \in E$ such that $p \in e$ and hence some $f' \in F(G)$, distinct from f, such that $e \subseteq \mathcal{F}(f')$. Since f' is disjoint from f^-, $p \in (\mathbb{S}^2 \setminus f^-)^- = \mathbb{S}^2 \setminus (f^-)^0$, i.e. $p \notin (f^-)^0$. Thus, the open set f satisfies $(f^-)^0 \subseteq f$, and so is regular open. By Lemma 2.4 (ii), $\bigcup F(G) \subseteq \sum F(G) \subseteq (\bigcup F(G))^- = \bigcup\{f^- \mid f \in F(G)\} = \mathbb{S}^2$. But by Lemma 2.3, $\sum F(G)$ is the unique regular open set lying between $\bigcup F(G)$ and its closure; i.e. $\sum F(G) = 1$. Hence, the elements of $F(G)$ form a connected partition in $\mathrm{RO}(\mathbb{S}^2)$. The last part of the lemma then follows from Lemma 2.3, since, if $f \in F(G)$, then both f and $\sum\{f' \in F(G') \mid f' \subseteq f\}$ are regular open sets sandwiched between $\bigcup\{f' \in F(G') \mid f' \subseteq f\}$ and its closure. QED

Of course, the converse of Lemma 2.79 is false: the configuration of Example 2.18 shows that not every connected partition in $\mathrm{RO}(\mathbb{S}^2)$ is the set of faces of some plane graph.

LEMMA 2.80 *If G is a piecewise linear plane graph such that G has no isolated vertices and every edge of G lies on the boundary of exactly 2 faces, then the faces of G form a connected partition in* $\mathrm{ROP}(\mathbb{S}^2)$.

Proof Let $L_1, \ldots, L_m$ be straight lines extending (in both directions) each of the line segments making up G. Let G' be the graph whose nodes are the points of intersection of the L_i (including ∞) and whose edges are the segments of the L_i joining them; and let P be the set of non-zero products $\pm s_1 \cdots \pm s_m$, where s_i is one of the residual half-planes of L_i for $1 \leq i \leq m$. By simple set-algebra, $\bigcup P = \bigcup F(G')$; and since every $r \in P$ is connected, and every $f \in F(G')$ is a maximal connected subset of $\mathbb{S}^2 \setminus |G'|$, $r \cap f \neq \emptyset$ implies $r \subseteq f$. Hence every $f \in F(G')$ is a union of elements of P. But since these elements are non-empty open and disjoint and f is connected, f simply is some element of P, and hence is an element of $\mathrm{ROP}(\mathbb{S}^2)$. Since $|G| \subseteq |G'|$, the result follows by the last part of Lemma 2.79. QED

Lemma 2.80 does have a converse:

LEMMA 2.81 *If $\bar{r}$ is a connected partition in* $\mathrm{ROP}(\mathbb{S}^2)$*, then $\bar{r}$ is the set of faces of some piecewise linear plane graph G; moreover, for any such plane graph G, G has no isolated vertices, and every edge of G lies on the boundary of exactly* 2 *faces.*

Proof By Claim 2.63, refine $\bar{r} = r_1, \ldots, r_n$ to a c^3-partition $\bar{t} = \{t_1, \ldots, t_N\}$, and let G_0 be the partition graph of $\bar{t}$. Suppose, by renumbering if necessary, that $r_1 = t_1 + \cdots + t_m$. Note that, if $e \in E(G_0)$, we have, for all j $(1 \leq j \leq N)$, $(e) \subseteq t_j^-$ or $(e) \cap t_j^- = \emptyset$. Hence if $(e) \not\subseteq r_1$, then $(e) \cap \bigcup_{m<j\leq N} t_j^- \neq \emptyset$, whence $e \subseteq t_j^-$ for some j $(m < j \leq N)$.

Let G_1 be the graph obtained from G_0 by first removing any edge e such that $(e) \subseteq r_1$, and then removing any vertex v such that $v \in r_1$. Since r_1 is open, the endpoints of every remaining arc are among the remaining vertices, so G_1 really is a plane graph. Moreover, if $m < j \leq N$, then $t_j^- \cap r_1 = \emptyset$, so that none of the vertices and edges removed from G_0 intersects t_j^-; hence t_j is a face of G_1. Therefore, the set of points

$$S = \{t_j \in F(G_0) \mid 1 \leq j \leq m\} \cup \\ \{e \in E(G_0) \mid (e) \subseteq r_1\} \cup \{v \in V(G_0) \mid v \in r_1\}$$

must be the union of some faces of G_1. Trivially, $S \subseteq r_1$. We claim that $r_1 \subseteq S$. For if $p \in \mathbb{S}^2$, exactly one of the following three cases holds: (i) $p \in t_j$ for some j; (ii) $p \in V(G_0)$; or (iii) $p \in e$ for some $e \in E(G_0)$. In case (i), either $p \in S$ or $p \notin r_1$, according as $j \leq m$. In case (ii), trivially, either $p \in S$ or $p \notin r_1$. In case (iii), if $p \notin S$, then $p \in (e) \not\subseteq r_1$, whence $p \in e \subseteq t_j^-$ for some j $(m < j \leq N)$, whence $p \in (\sum_{m<j<N} t_j)^- = \mathbb{S}^2 \setminus r_1$. This proves that $r_1 \subseteq S$. Thus, $r_1 = S$ is the union of a number of faces of G_1. But r_1 is by assumption connected, so r_1 is a face of G_1. Proceeding in the same way for $r_2, \ldots, r_n$, we obtain the desired graph $G = G_n$. QED

Lemmas 2.80 and 2.81 concern the mereotopology $\mathrm{ROP}(\mathbb{S}^2)$, but almost exactly similar arguments can be given for $\mathrm{ROS}(\mathbb{S}^2)$ and $\mathrm{ROQ}(\mathbb{S}^2)$. We omit the details, which are routine. Summarizing, we have:

THEOREM 2.82 *A tuple $\bar{u}$ of subsets of $\mathbb{S}^2$ is a connected partition in* $\mathrm{ROS}(\mathbb{S}^2)$ (*alternatively:* $\mathrm{ROP}(\mathbb{S}^2)$, $\mathrm{ROQ}(\mathbb{S}^2)$) *if and only if it is the set of faces of a semi-algebraic* (*respectively: piecewise linear, rational piecewise linear*) *graph with no isolated vertices and every edge lying on the boundary of two faces.*

We now have at our disposal the means to give the promised proof of the only-if direction of Lemma 2.49. We therefore digress from the main business of this section to do so.

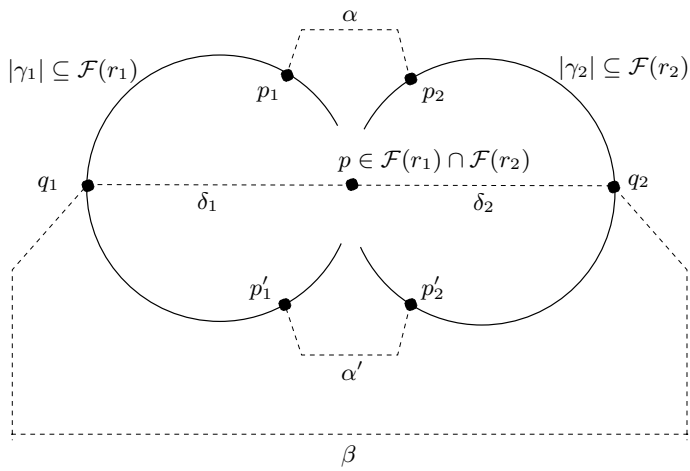


Figure 2.12. Proof of Lemma 2.83.

LEMMA 2.83 *Let M be a finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection, and let r_1, r_2 be elements of M such that $r_1^- \cap r_2^- \neq \emptyset$. Then there exist elements r'_1, r'_2, t of M such that: (i) $r'_1 \leq r_1$ and $r'_2 \leq r_2$; (ii) $-(r'_1 + t)$, $-(r'_2 + t)$ and t are all connected; (iii) $-(r'_1 + r'_2 + t)$ is not connected.*

Proof We may suppose without loss of generality that r_1 and r_2 are elements in a c^3-partition and that $r_1 + r_2 \neq 1$. For otherwise, by Claims 2.54 and 2.63, we can easily find a c^3-partition consisting of at least three elements, and containing distinct elements $\hat{r}_i \leq r_i$ $(i = 1, 2)$ such that $\hat{r}_1^- \cap \hat{r}_2^- \neq \emptyset$; now proceed with r_i replaced by $\hat{r}_i$.

Let $s = -(r_1 + r_2)$. From the properties of c^3-partitions, s, $r_1 + s$ and $r_2 + s$ are all connected. Moreover, s is the sum of various cells of the c^3-partition, whence, by Claim 2.61, each r_i $(i = 1, 2)$ forms a connected sum with at least one of these cells, s_i, say. Now, by Lemma 2.73, $\mathcal{F}(r_i) \cap \mathcal{F}(s_i)$ is connected, and is certainly not empty or a singleton (by the connectedness of $r_i + s_i$), so that $\mathcal{F}(r_i) \cap \mathcal{F}(s_i)$ includes the locus of some Jordan arc γ_i. (Note that $|\gamma_i| \subseteq \mathcal{F}(r_i) \cap \mathcal{F}(s)$.) Now choose internal points p_i, q_i and p'_i of γ_i, for $i = 1, 2$, in the order shown in Fig. 2.12. Since M has curve-selection, let α be a cross-cut in s from p_1 to p_2, α' a cross-cut in s from p'_1 to p'_2, and β a cross-cut in s from q_1 to q_2, drawn as shown. Since $|\beta|$ is compact, we can cover it with finitely many small elements of M whose closures are disjoint from $|\alpha| \cup |\alpha'|$. Taking the sum of these elements to be t'', let t' be the component of $t'' \cdot s$ which includes $(|\beta| \setminus \{p, q\})$. Note that, by Lemma 2.13, we may freely take components of elements of M.

Since M is finitely decomposable, consider the components $d_1, \ldots, d_k$ of $-(r_1 + r_2 + t') = s \cdot -t'$. Since the sets t', $d_1, \ldots, d_k$ are all connected, and their sum is s, which is also connected, Claim 2.61 implies that $t' + d_j$ is connected for all j ($1 \leq j \leq k$). Since $r_1^- \cap r_2^-$ contains a point p, and M has curve selection, let δ_i be a cross-cut in r_i from p to q_i. Then $|\delta_1| \cup |\delta_2| \cup |\beta|$ separates $|\alpha|$ and $|\alpha'|$ (see Fig. 2.12). But $|\delta_1| \cup |\delta_2| \cup |\beta| \subseteq (r_1 + r_2 + t')^-$; that is, α and α' lie in different components of $-(r_1 + r_2 + t')$. Without loss of generality, suppose α lies in d_1 and α' in d_2. Now let $t = t' + d_3 + \cdots + d_k$.

Since all the $t' + d_j$ are connected, t is connected. Moreover, it is easy to see that $p_i \in d_1 + r_i$ and $p_i' \in d_2 + r_i$ ($i = 1, 2$), whence the four sets $d_j + r_i$ ($i = 1, 2$; $j = 1, 2$) are all connected. It follows that $-(t + r_1) = r_2 + d_1 + d_2$ and $-(t + r_2) = r_1 + d_1 + d_2$ are both connected. On the other hand, $-(t + r_1 + r_2) = d_1 + d_2$ is not connected. Setting $r_1' = r_1$ and $r_2' = r_2$ gives us the desired elements r_1', r_2', t. QED

4.4 Expressive power of first-order languages in plane mereotopologies

We are now in a position to give an absolute characterization of the expressive power of the languages $L_{c,\leq}$ and L_C over certain mereotopologies of interest. Recall the concept of topologically complete formula given in Definition 2.51. The following notation will be useful in constructing topologically complete formulas.

NOTATION 2.84 *Given a fixed Boolean algebra, a* Boolean matrix *is a rectangular matrix whose entries are the elements* 1 *and* 0. *If $\bar{r}$ is an n-tuple, $\bar{s}$ an N-tuple, and A a Boolean matrix with N rows and n columns, we write $\bar{r} = \bar{s}A$ to indicate that each element of $\bar{r}$ is the sum of certain elements of $\bar{s}$ as indicated by the elements of A via normal matrix multiplication. Similarly, we write $\bar{x} = \bar{z}A$ in first-order formulas to abbreviate the obvious conjunction of Boolean algebra equations.*

THEOREM 2.85 *Let M be any finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection, and let Σ be the signature $(c, \leq, +, \cdot, -)$. Every tuple from M satisfies some (purely existential) L_Σ-formula which is topologically complete in M over $\mathbb{S}^2$.*

Proof Writing $\bar{z}$ for $z_1, \ldots, z_N$, let $\psi_{c^3}^N(\bar{z})$ be the formula:

$$\bigwedge_{1 \leq i \leq N} (c(z_i) \wedge z_i > 0) \wedge \bigwedge_{1 \leq i \leq j \leq N} (c(-(z_i + z_j)) \wedge z_i \cdot z_j = 0) \wedge \sum_{1 \leq i \leq N} z_i = 1.$$

Thus, $M \models \psi^N_{\mathrm{c}^3}[\bar{s}]$ if and only if $\bar{s}$ is an N-element c^3-partition. If $\bar{s} = s_1, \ldots, s_N$ is a c^3-partition in M, let $\psi^{\bar{s}}_+(\bar{z})$ be the formula:

$$\bigwedge\{c(z_i + z_j) \mid 1 \leq i < j \leq N \text{ and } s_i + s_j \text{ is connected}\} \wedge$$
$$\bigwedge\{\neg c(z_i + z_j) \mid 1 \leq i < j \leq N \text{ and } s_i + s_j \text{ is not connected}\},$$

where $\bar{z}$ is the tuple of variables $z_1, \ldots, z_n$. Thus, $\psi^{\bar{s}}_+(\bar{z})$ encodes the neighbourhood structure of $\bar{s}$. Now let $\bar{r} = r_1, \ldots, r_n$ be any tuple of elements of M. By Claim 2.63, there exists a c^3-partition $\bar{s} = s_1, \ldots, s_N$ in M and a Boolean matrix A such that $\bar{r} = \bar{s}A$. Writing $\bar{x}$ for $x_1, \ldots, x_n$, let $\psi_{\bar{r}}(\bar{x})$ be the formula

$$\exists \bar{z}(\psi^N_{\mathrm{c}^3}(\bar{z}) \wedge \psi^{\bar{s}}_+(\bar{z}) \wedge \bar{x} = \bar{z}A).$$

Certainly, $M \models \psi_{\bar{r}}[\bar{r}]$. And if $\bar{r}'$ is a tuple from M such that $M \models \psi_{\bar{r}}[\bar{r}']$, let $\bar{s}' = s'_1, \ldots, s'_N$ be corresponding witnesses for the existentially quantified variables $\bar{z}$. Then $s_1, \ldots, s_N$ and $s'_1, \ldots, s'_N$ are c^3-partitions in $\mathbb{S}^2$ which have the same neighbourhood structure, and hence which are similarly situated in $\mathbb{S}^2$, by Theorem 2.78. It follows that $\bar{r}$ and $\bar{r}'$ are similarly situated in $\mathbb{S}^2$ too. Thus, $\psi_{\bar{r}}(\bar{x})$ is topologically complete in M over $\mathbb{S}^2$. QED

COROLLARY 2.86 *Let M be any finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection. Every tuple from M satisfies some L_C-formula which is topologically complete in M over $\mathbb{S}^2$.*

Proof Theorem 2.85 and Lemmas 2.22, 2.27 and 2.49. QED

Thus, for certain well-behaved mereotopologies over $\mathbb{S}^2$, both L_C and $L_{c,\leq}$ are, as we might put it, "topologically fully descriptive".

We now turn to the question of expressive power in mereotopologies over $\mathbb{R}^2$. We need some auxiliary lemmas.

LEMMA 2.87 *Let $\bar{r} = r_1, \ldots, r_n$ be a c^3-partition in any mereotopology M over $\mathbb{S}^2$ having curve-selection. Let $p, p' \in \mathbb{S}^2$ such that, for all i $(1 \leq i \leq n)$, $p \in r_i^-$ if and only if $p' \in r_i^-$. Then there is a homeomorphism $h : \mathbb{S}^2 \to \mathbb{S}^2$ mapping p to p' and fixing each r_i.*

Proof Obvious, by viewing $\bar{r}$ as $F(G)$ for some plane graph G. QED

LEMMA 2.88 *Let $\bar{r} = r_1, \ldots, r_n$ and $\bar{r}' = r'_1, \ldots, r'_n$ be similarly situated c^3-partitions in any mereotopology M over $\mathbb{S}^2$ having curve-selection. Let $p \in \mathbb{S}^2$ be such that, for all i $(1 \leq i \leq n)$, $p \in r_i^-$ if and only if $p \in r_i'^-$. Then there is a homeomorphism $h : \mathbb{S}^2 \to \mathbb{S}^2$ fixing p and mapping $\bar{r}$ to $\bar{r}'$.*

Proof Let $h' : \mathbb{S}^2 \to \mathbb{S}^2$ be some homeomorphism mapping $\bar{r}$ to $\bar{r}'$. Then, for all i ($1 \leq i \leq n$), $h'(p) \in {r'_i}^-$ if and only if $p \in {r_i}^-$. By Lemma 2.87, let $h'' : \mathbb{S}^2 \to \mathbb{S}^2$ be a homeomorphism fixing each r'_i, and mapping $h'(p)$ to p. Then $h := h'' \circ h'$ has the required properties. QED

THEOREM 2.89 *Let M be any finitely decomposable mereotopology over $\mathbb{R}^2$ such that $\dot{M}$ has curve-selection. Every tuple from M satisfies some L_C-formula which is topologically complete in M over $\mathbb{R}^2$.*

Proof Given any tuple $s_1, \ldots, s_N$ from M, let $\phi^{\bar{s}}_{\infty}(\bar{z})$ be the L_C-formula

$$\bigwedge\{\phi_{\mathrm{b}^2}(z_i) \mid 1 \leq i \leq N \text{ and } s_i \text{ is bounded}\} \wedge \bigwedge\{\neg\phi_{\mathrm{b}^2}(z_i) \mid 1 \leq i \leq N \text{ and } s_i \text{ is not bounded}\},$$

where $\bar{z}$ is the tuple of variables $z_1, \ldots, z_N$, and ϕ_{b^2} is as in Lemma 2.30. Thus, $\phi^{\bar{s}}_{\infty}(\bar{z})$ encodes the pattern of boundedness in the tuple $\bar{s}$. Now, given a tuple $\bar{r}$, let $\bar{s}$ be an N-element c^3-partition in M refining $\bar{r}$, and let A be a Boolean matrix satisfying $\bar{r} = \bar{s}A$. Using the translation from $L_{c,\leq}$ to L_C established by Lemmas 2.22 and 2.27, let $\phi^N_{\mathrm{c}^3}(\bar{z})$ and $\phi^{\bar{s}}_{+}(\bar{z})$ be the L_C-formulas corresponding to $\psi^N_{\mathrm{c}^3}(\bar{z})$ and $\psi^{\bar{s}}_{+}(\bar{z})$ in the proof of Theorem 2.85. Writing $\bar{x}$ for $x_1, \ldots, x_n$, let $\phi_{\bar{r}}(\bar{x})$ be the formula

$$\exists\bar{z}(\phi^N_{\mathrm{c}^3}(\bar{z}) \wedge \phi^{\bar{s}}_{+}(\bar{z}) \wedge \phi^{\bar{s}}_{\infty}(\bar{z}) \wedge \bar{x} = \bar{z}A).$$

Certainly, $M \models \psi_{\bar{r}}[\bar{r}]$; and if $\bar{r}'$ is a tuple from M such that $M \models \psi_{\bar{r}}[\bar{r}']$, let $\bar{s}' = s'_1, \ldots, s'_N$ again be a corresponding witnesses for the existentially quantified variables $\bar{z}$. Then $\dot{s}_1, \ldots, \dot{s}_N$ and $\dot{s}'_1, \ldots, \dot{s}'_N$ are c^3-partitions in $\mathbb{S}^2$ which have the same neighbourhood structure, so that by Theorem 2.78 and Lemma 2.88, there is a homeomorphism $h : \mathbb{S}^2 \to \mathbb{S}^2$ fixing ∞ and mapping each $\dot{s}_i$ to $\dot{s}'_i$. Hence $\bar{s}$ and $\bar{s}'$ are similarly situated in $\mathbb{R}^2$, whence $\bar{r}$ and $\bar{r}'$ are similarly situated in $\mathbb{R}^2$ too. QED

Thus, for well-behaved mereotopologies over $\mathbb{R}^2$, L_C is, as we might put it, "topologically fully descriptive".

4.5 Homogeneous mereotopologies

Up to this point, we have been concerned only to show that certain relations *can* be defined by first-order formulas with signatures of topological primitives. We turn now briefly to the question of which relations *cannot* be so defined.

At first glance, one might assume that languages with purely topological primitives can express only topological concepts in mereotopologies over which they are interpreted. However, this assumption is correct only if the

mereotopologies in question have a certain property. Recall that, for a fixed topological space X, we write $\bar{u} \sim \bar{v}$ to mean that the tuples of subsets $\bar{u}$ and $\bar{v}$ are similarly situated in X (Definition 2.51).

DEFINITION 2.90 *Let M be a mereotopology over X. We say M is* homogeneous (over X) *if, given any tuples $\bar{r}, \bar{s}$ from M with $\bar{r} \sim \bar{s}$ and any element $r \in M$, there exists an element $s \in M$ with $\bar{r}, r \sim \bar{s}, s$. Let M' also be a mereotopology over X, with $M' \subseteq M$. We say M' is* homogeneously embedded in M (*over X*) *if, given any tuple $\bar{r}$ from M', and any $r \in M$, there exists $s \in M'$ with $\bar{r}, r \sim \bar{r}, s$.*

LEMMA 2.91 *Let X be either $\mathbb{R}^2$ or $\mathbb{S}^2$, and let M be any of* $\mathrm{ROS}(X)$, $\mathrm{ROP}(X)$ *or* $\mathrm{ROQ}(X)$. *Then M is homogeneous.*

Proof Assume $M = \mathrm{ROS}(\mathbb{S}^2)$; the other cases are identical. Let $\bar{r}$, $\bar{s}$ be tuples from M, and let $r \in M$. Let $\bar{t}$ be a connected partition refining $\bar{r}, r$, which by Theorem 2.82 is the set of faces of some semi-algebraic plane graph G. If $h : \mathbb{S}^2 \to \mathbb{S}^2$ is a homeomorphism mapping $\bar{r}$ to $\bar{s}$, then h maps G to a plane graph H. But then it is not difficult to show that the edges of H can be deformed into a semi-algebraic plane graph H', and moreover, that this may be done in such a way that existing semi-algebraic edges are unaffected. By Theorem 2.82, the faces of the resulting graph are elements of M; hence we have a homeomorphism mapping $\bar{r}$ to $\bar{s}$ and taking r to some element s of M. QED

Homogeneity and homogeneous embedding are important because of the following facts.

LEMMA 2.92 *Let M be a homogeneous mereotopology over a topological space X, and fix a signature Σ of topological primitives. If $\bar{r}$ and $\bar{s}$ are tuples of M which are similarly situated in X, then $\bar{r}$ and $\bar{s}$ satisfy the same L_Σ-formulas in M.*

Proof We show by induction on the complexity of $\phi(\bar{x}) \in L_\Sigma$ that, if $\bar{r}$ and $\bar{s}$ are tuples of the appropriate arity which are similarly situated in X, then $M \models \phi[\bar{r}]$ implies $M \models \phi[\bar{s}]$. The base case follows from the fact that the primitives in Σ have topological interpretations. The only non-trivial recursive case is where $\phi[\bar{x}] = \exists y \psi(\bar{x}, y)$. If $M \models \phi[\bar{r}]$, there exists $r \in M$ such that $M \models \psi[\bar{r}, r]$, and by homogeneity, if $\bar{r} \sim \bar{s}$, there exists $s \in M$ such that $\bar{r}, r \sim \bar{s}, s$, whence $M \models \psi[\bar{s}, s]$ by inductive hypothesis, so that $M \models \phi[\bar{s}]$ as required. QED

Lemma 2.92 gives an upper bound on the expressive power of first-order languages with signatures of topological primitives interpreted over homogeneous mereotopologies: such languages cannot distinguish between similarly situated

tuples. It thus provides a partial converse to Theorems 2.85 and 2.89. It also yields an easy proof that, over well-behaved open-plane mereotopologies, $L_{c,\leq}$ cannot express the property of being bounded:

THEOREM 2.93 *Let M be a mereotopology over $\mathbb{R}^2$ such that $\dot{M}$ is homogeneous, and suppose M has curve-selection and contains a region r similarly situated in $\mathbb{R}^2$ to the open unit disc $\mathbf{B}^2$. Then there exists no formula $\psi(x)$ of $L_{c,\leq}$ such that, for all $r \in M$, r is bounded if and only if $M \models \psi[r]$.*

Proof Suppose such a formula $\psi(x)$ exists. Then $M \models \psi[r]$, and by Lemma 2.38, $\dot{M} \models \psi[\dot{r}]$. Since M has curve-selection, by Proposition 2.44 both $\dot{r}$ and its complement $-(\dot{r})$ in $\dot{M}$ are 2-cells in $\mathbb{S}^2$, and hence are similarly situated. By Lemma 2.92, $\dot{M} \models \psi[-(\dot{r})]$, and so by Lemma 2.38, $M \models \psi[-r]$. This contradicts the fact that $-r$ is unbounded. QED

Finally, we return to the relationship between $\mathrm{ROS}(X)$, $\mathrm{ROP}(X)$ and $\mathrm{ROQ}(X)$.

LEMMA 2.94 *Let X be either $\mathbb{R}^2$ or $\mathbb{S}^2$. Then $\mathrm{ROQ}(X)$ is homogeneously embedded in $\mathrm{ROP}(X)$, which is in turn homogeneously embedded in $\mathrm{ROS}(X)$.*

Proof Virtually identical to the proof of Lemma 2.91. QED

The following result is well-known (see, for example, Hodges, 1993 p. 55).

PROPOSITION 2.95 (TARSKI-VAUGHT) *Let $\mathfrak{A}$, $\mathfrak{B}$ be structures with $\mathfrak{A} \subseteq \mathfrak{B}$, and suppose that, for any n-tuple $\bar{a}$ from A and any formula $\phi(\bar{x})$ of the form $\exists y \psi(\bar{x}, y)$ such that $\mathfrak{B} \models \phi[\bar{a}]$, there exists $a \in A$ such that $\mathfrak{B} \models \psi[\bar{a}, a]$. Then $\mathfrak{A} \preceq \mathfrak{B}$.*

LEMMA 2.96 *Let M, M' be mereotopologies over a topological space X, with M homogeneous and M' homogeneously embedded in M. Fix a signature of topological primitives. Then $M' \preceq M$.*

Proof By assumption, $M' \subseteq M$. Let $\bar{r}$ be an n-tuple of elements of M', and let $\phi(\bar{x})$ be any formula of the form $\exists y \psi(\bar{x}, y)$ such that $M \models \phi[\bar{r}]$. Then there exists $r \in M$ such that $M \models \psi[\bar{r}, r]$. Since M' is homogeneously embedded in M, there exists $s \in M'$ such that $\bar{r}, r \sim \bar{r}, s$. Since M is homogeneous, $M \models \psi[\bar{r}, s]$ by Lemma 2.92. The result then follows by Proposition 2.95. QED

Hence, for X either $\mathbb{R}^2$ or $\mathbb{S}^2$, and over any signature Σ of topological primitives, we have $\mathrm{ROQ}(X) \preceq \mathrm{ROP}(X) \preceq \mathrm{ROS}(X)$. In particular, these three structures have identical L_Σ-theories. We show in the sequel that this is no accident: almost any 'reasonable' mereotopology over $\mathbb{S}^2$ has the same

L_Σ-theory. Anticipating these results, we employ the following notation and terminology.

DEFINITION 2.97 *Let Σ be a signature of topological primitives. We call the theory* $\mathrm{Th}_\Sigma(\mathrm{ROS}(\mathbb{S}^2))$ *the* standard L_Σ-theory (of closed plane mereotopology)*, and denote it* $\mathbf{T}_\Sigma$.

5. Axiomatization

In this section, we provide an axiomatic characterization of $\mathbf{T}_{c,\leq}$, the standard $L_{c,\leq}$-theory of closed plane mereotopology. The material is essentially that of Pratt and Schoop, 1998. The axiom system in question will help us to identify mereotopologies over $\mathbb{S}^2$ having the standard $L_{c,\leq}$-theory.

As before, we write $\psi^n_{\mathrm{c}^3}(\bar{x})$ for the $L_{c,\leq}$-formula stating that $\bar{x}$ forms an n-element c^3-partition, and $x = u \oplus v$ for the $L_{c,\leq}$-formula stating that u and v are disjoint, non-zero, connected regions summing to x. Let M be a mereotopology over $\mathbb{S}^2$ having curve-selection. Consider a triple r, s, t from M satisfying the formula $\psi^3_{\mathrm{c}^3}(x, y, z)$. By Lemma 2.45, each of these regions is a 2-cell, and it is easy to see that the closures of any two of these intersect in a Jordan arc. (Formally, this follows by Lemma 2.73.) Now let ψ_{split} denote the $L_{c,\leq}$-formula

$$\forall x \forall y \forall z \big(\psi^3_{\mathrm{c}^3}(x, y, z) \rightarrow \\ \exists u \exists v \big(u \oplus v = x \wedge c(u + y) \wedge \neg c(u + z) \wedge c(v + z) \wedge \neg c(v + y)\big)\big).$$

Informally, ψ_{split} "says" that, given two 2-cells r and s whose frontiers intersect in a Jordan arc, r can be partitioned into two connected regions using a cross-cut whose end-points are the end-points of that Jordan arc (Fig. 2.13a).

DEFINITION 2.98 *A mereotopology M is* splittable *if $M \models \psi_{\mathrm{split}}$.*

The following lemma is unsurprising.

LEMMA 2.99 *The mereotopologies* $\mathrm{ROS}(\mathbb{S}^2)$*,* $\mathrm{ROP}(\mathbb{S}^2)$ *and* $\mathrm{ROQ}(\mathbb{S}^2)$ *are splittable.*

Proof Almost immediate from Theorem 2.85, Lemma 2.42 and Proposition 2.43. QED

However, not all finitely decomposable mereotopologies over $\mathbb{S}^2$ having curve-selection are splittable. If an $(n-1)$-dimensional hyperplane in $\mathbb{R}^n$ is defined by an equation $x_i = 0$, where $1 \leq i \leq n$, we call it an *axis-oriented hyperplane*; and if a half-space is bounded by an axis-oriented hyperplane, we call it an *axis-oriented half-space*.

EXAMPLE 2.100 *Define* $\mathrm{ROX}(\mathbb{S}^n)$ *to be the Boolean sub-algebra of* $\mathrm{RO}(\mathbb{S}^n)$ *generated by the axis-oriented half-spaces. It is easy to see that* $\mathrm{ROX}(\mathbb{S}^n)$ *is a*

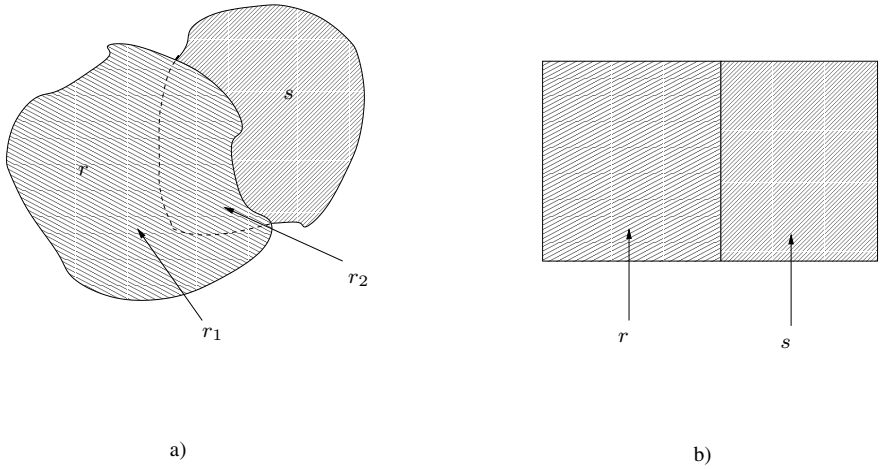


Figure 2.13. a) The configuration of ψ_{split}: r and s are disjoint 2-cells with $r^- \cap s^-$ a Jordan arc; r is broken into r_1 and r_2. b) A pair of regions in $\text{ROX}(\mathbb{R}^2)$ violating ψ_{split}.

finitely decomposable mereotopology over $\mathbb{S}^n$ having curve-selection. However, $\text{ROX}(\mathbb{S}^n) \not\models \psi_{\text{split}}$, as is clear in the case $n = 2$ by inspection of Fig. 2.13b).

Thus, whereas $\text{RO}(\mathbb{S}^2)$ has, as it were, too many regions for the standard theory, $\text{ROX}(\mathbb{S}^2)$ has too few. As we have observed, $\text{RO}(\mathbb{S}^2)$ is not finitely decomposable, and lacks curve-selection, while $\text{ROX}(\mathbb{S}^2)$ is not splittable. It transpires that these represent the *only* ways of failing to exhibit the standard theory of closed plane mereotopology. Specifically, we show in this section that all splittable, finitely decomposable mereotopologies over $\mathbb{S}^2$ having curve-selection have the same $L_{c,\leq}$-theory. Our strategy is to pick one splittable, finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection—$\text{ROP}(\mathbb{S}^2)$ will do—and characterize its theory axiomatically. We then merely need to check that our axiom system is correct for all such mereotopologies.

5.1 The axioms

Our axiom system comprises three parts: a *general inference system*, a set of *proper axioms* and an *ω-rule*. (i) The general inference system is simply any complete Hilbert system for first-order logic, restricted to the signature $\{+, \cdot, -, \leq, c\}$. (ii) The proper axioms are as follows:

1 the usual axioms of Boolean algebra, and the axiom $0 \neq 1$;

2 the axiom ψ_{con} (Lemma 2.53);

3 where $n > 2$, the axioms ψ^n_{sum} (Lemma 2.56);

4 the axiom

$$\neg\exists x_1 \ldots \exists x_5 \Big(\bigwedge_{1\leq i\leq 5} (c(x_i) \wedge x_i \neq 0) \wedge \bigwedge_{1\leq i<j\leq 5} (c(x_i + x_j) \wedge x_i \cdot x_j = 0)\Big);$$

5 the axiom

$$\neg\exists x_1 \ldots \exists x_6 \Big(\bigwedge_{1\leq i\leq 6} (c(x_i) \wedge x_i \neq 0) \wedge \bigwedge_{1\leq i<j\leq 6} x_i \cdot x_j = 0 \wedge \bigwedge_{\substack{1\leq i\leq 3\\ 4\leq j\leq 6}} c(x_i + x_j)\Big);$$

6 the axioms $c(0)$ and $c(1)$;

7 the axiom ϕ_{break} (Lemma 2.57);

8 the axiom ϕ_{split} (Definition 2.98).

(iii) The final component of our axiom system is the ω-rule. If $n \geq 1$, we let $\psi^n_{\text{c}}(x)$ stand for the formula

$$\exists z_1 \ldots \exists z_n \Big(\bigwedge_{1\leq i\leq n} c(z_i) \wedge (x = z_1 + \cdots + z_n)\Big).$$

Thus, $\psi^n_{\text{c}}(x)$ "says" that x can be formed by summing n connected regions. The ω-rule is then the (infinitary) rule of inference:

$$\frac{\{\forall x(\psi^n_{\text{c}}(x) \to \phi(x)) | n \geq 1\}}{\forall x \phi(x)}.$$

Let Φ be a set of $L_{c,\leq}$-sentences. A *proof* with *premises* Φ in the above system is a sequence of $L_{c,\leq}$-formulas $\{\phi_\alpha\}_{\alpha<\beta}$, for some ordinal β (not necessarily finite) such that every ϕ_α is either (i) an element of Φ or (ii) an axiom or (iii) the result of applying a rule of inference to some formulas ϕ_γ with $\gamma < \alpha$. If ψ is the last line of some such proof, we write $\Phi \vdash \psi$. If $\Phi = \{\phi\}$ we write $\phi \vdash \psi$, and if $\Phi = \emptyset$ we write $\vdash \psi$ and call ψ a *theorem*. Let us denote the set of theorems by T_{Ax}. The main result of this section is:

THEOREM 2.101 *T_{Ax} is the complete $L_{c,\leq}$-theory of any finitely decomposable, splittable mereotopology over $\mathbb{S}^2$ having curve-selection.*

Proof Lemmas 2.103 and 2.105, below. QED

Of course, this entails that all such mereotopologies, considered as $\{c, \leq\}$-structures are elementarily equivalent.

The ω-rule is less unfamiliar than one might at first think. Essentially, it says that if a property holds of every region which is the sum of finitely many connected regions, then it simply holds of every region. This conditional is obviously true in a finitely decomposable mereotopology. Thus, a proof involving the ω-rule is analogous to an argument of the kind encountered in elementary algebra textbooks in which one proves a property of all polynomials by showing that it holds of all polynomials of some arbitrary degree n. Nevertheless, the inclusion of an infinitary proof rule does mean that we ought to check the deduction theorem.

LEMMA 2.102 *Let ϕ be an $L_{c,\leq}$-sentence and ψ an $L_{c,\leq}$-formula such that $\phi \vdash \psi$. Then $\vdash \phi \to \psi$.*

Proof By assumption, there is a proof $\{\phi_\alpha\}_{\alpha<\beta+1}$ with premises $\{\phi\}$ and last line $\phi_\beta = \psi$. Without loss of generality, we may assume that the first (actually, zeroth) line of the proof ψ_0 is ϕ. We proceed by induction on β. The case $\beta = 0$ is trivial, since $\vdash \phi \to \phi$. If $\beta > 0$, then either ϕ_β is an axiom or is derived from applying a rule of inference to earlier lines of the proof. The only interesting case is where $\phi = \forall x \pi$ is derived by the ω-rule from the formulas $\forall x(\psi_{\mathrm{c}}^n(x) \to \pi)$ occurring earlier in the proof. But the inductive hypothesis then yields $\vdash \phi \to \forall x(\psi_{\mathrm{c}}^n(x) \to \pi)$, for each n, whence $\vdash \forall x(\psi_{\mathrm{c}}^n(x) \to (\phi \to \pi))$. The ω-rule then yields $\vdash \forall x(\phi \to \pi)$, whence $\vdash \phi \to \forall x \pi$ (note that ϕ is a sentence), as required. QED

We remark in passing that the axiom $c(0)$ is actually redundant: it can be derived from the other axioms and proof rules.

5.2 Correctness

In this section, we establish the easy half of Theorem 2.101.

LEMMA 2.103 *If M is a splittable, finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection, then $M \models T_{\mathrm{Ax}}$.*

Proof We follow the enumeration in Sec. 5.1, showing that the proper axioms are all true in M and that the ω-rule is truth-preserving.

1 M is a mereotopology.

2 Lemma 2.53.

3 Lemma 2.56.

4 Suppose $r_1, \ldots, r_5$ are connected, non-empty and pairwise disjoint, and that any pair of them have a connected sum. By Lemma 2.26, choose points $p_i \in r_i$ and $q_{i,j} \in \mathcal{F}(r_i) \cap \mathcal{F}(r_j) \cap (r_i + r_j)$ $(1 \leq i < j \leq 5)$. For each i $(1 \leq i \leq 5)$, draw end-cuts in r_i from p_i to all the points $q_{i,j}$ and $q_{j,i}$; it is easy to see that these can be chosen so that any pair of these end-cuts intersect only in the point p_i. Ignoring the points $q_{i,j}$, we have a plane drawing of the graph K^5, which is known to be non-planar (Fig. 2.10).

5 As for axiom 4, but with $K_{3,3}$ instead of K^5.

6 Trivial.

7 Lemma 2.57.

8 M is splittable.

The ω-rule is obviously truth-preserving, because M is finitely decomposable.

QED

5.3 Completeness

In this section, we establish the difficult half of Theorem 2.101. We make use of the *omitting types theorem*: for details, see, e.g. Hodges, 1993, pp 333. Let $\mathfrak{A}$ be a structure, $\Phi(x)$ a set of formulas with free variable x, and T a set of sentences. We say that $\mathfrak{A}$ *omits* $\Phi(x)$ if, for all $a \in A$, $\mathfrak{A} \not\models \Phi[a]$. We say that T *locally omits* $\Phi(x)$ if, for every formula $\theta(x)$ with free variable x such that θ is consistent with T, there exists $\phi(x) \in \Phi(x)$ such that $T \not\models \forall x(\theta(x) \rightarrow \phi(x))$. The following theorem is a well-known strengthening of the completeness theorem for first-order logic.

PROPOSITION 2.104 (OMITTING TYPES THEOREM) *If a consistent theory T locally omits a set of formulas $\Phi(x)$, then T has a countable model omitting $\Phi(x)$.*

With these preliminaries behind us, we can proceed with our completeness proof.

LEMMA 2.105 *If ϕ is an $L_{c,\leq}$-sentence, and* $\mathrm{ROP}(\mathbb{S}^2) \models \phi$, *then* $\phi \in T_{\mathrm{Ax}}$.

Proof Suppose that $\phi \notin T_{\mathrm{Ax}}$. We are required to prove that $\mathrm{ROP}(\mathbb{S}^2) \models \neg\phi$. Let T be the set of all and only those $L_{c,\leq}$-sentences ψ such that $\neg\phi \vdash \psi$. By Lemma 2.102, T is a consistent set of sentences, and from the ω-rule, T locally omits the type $\{\neg\psi_{\mathrm{c}}^n(x) \mid n > 0\}$. By Proposition 2.104, there exists a countable model $\mathfrak{A} \models T$ omitting that type. Fix the structure $\mathfrak{A}$ for the remainder of this proof.

We now proceed in three stages. Stage 1 establishes some basic facts about $\mathfrak{A}$; Stage 2 shows that $\mathfrak{A}$ can be embedded in the $\{c, \leq\}$-structure $\mathrm{ROP}(\mathbb{S}^2)$; Stage 3 shows that the embedding we have chosen is in fact elementary.

Stage 1: Axioms 1 ensure that the reduct of $\mathfrak{A}$ to the signature $\{+, \cdot, -, \leq\}$ is a Boolean algebra. Such structures were discussed in Sec. 4.1, where various terminology and notational conventions were introduced. We carry these over to the present proof. Using that terminology, another way of saying that $\mathfrak{A}$ omits the type $\{\neg\psi_{\mathrm{c}}^n(x) \mid n > 0\}$ is to say that $\mathfrak{A}$ is finitely decomposable.

By Axioms 2, 3, 6 and 7, all the claims in Sections 4.1 and 4.2 hold of $\mathfrak{A}$. In particular, every tuple can be refined to a connected partition, and thence to a c^2- and a c^3-partition. Furthermore, we have

CLAIM 2.106 *Let $\bar{b} = b_1, \ldots, b_n$ be a connected partition in $\mathfrak{A}$. Then the neighbourhood graph of $\bar{b}$ is planar.*

Proof By Proposition 2.68, if the neighbourhood graph G of $\bar{b}$ is not planar, it contains either K^5 or $K_{3,3}$ as a minor. But then there is a sequence of contractions of G resulting in a graph H which has either K^5 or $K_{3,3}$ as a sub-graph. By repeated applications of Claim 2.60 (re-numbering the b_i as necessary), there is a connected partition $\bar{s}$ in $\mathfrak{A}$ whose neighbourhood graph contains K^5 or $K_{3,3}$ as a sub-graph. But this is impossible by Axioms 4 and 5.

QED

Stage 2: Since $\mathfrak{A}$ is countable, let $A = \{a_1, a_2, \ldots\}$. Let $N_0 = 1$ and let $\bar{c}^0$ be the 1-tuple whose element is the unit of the Boolean algebra $\mathfrak{A}$. Trivially, $\bar{c}^0$ is a c^3-partition. For $n \geq 0$, suppose that the c^3-partition $\bar{c}^{(n)} = c_1^{(n)}, \ldots, c_{N_n}^{(n)}$ in $\mathfrak{A}$ has been defined; then, by Claim 2.63, let $\bar{c}^{(n+1)} = c_1^{(n+1)}, \ldots, c_{N_{n+1}}^{(n+1)}$ be a c^3-partition in $\mathfrak{A}$ refining the tuple $c_1^{(n)}, \ldots, c_{N_n}^{(n)}, a_{n+1}$. It is then obvious that, for each $n > 0$, $\bar{c}^{(n)}$ refines the tuple $a_1, \ldots, a_n$ and also the tuple $\bar{c}^{(m)}$ for all m $(0 < m \leq n)$. We fix the enumerations $a_0, a_1, \ldots$ and $\bar{c}^{(0)}, \bar{c}^{(1)}, \ldots$ for the remainder of Stage 2.

For brevity, denote $\mathrm{ROP}(\mathbb{S}^2)$ by S. We now map each initial segment $a_1, \ldots, a_n$ of A into S. Let $w^{(n)}$ be the set of functions $g^{(n)} : \{c_1^{(n)}, \ldots, c_{N_n}^{(n)}\} \to S$ satisfying the conditions:

G1: the regions $g^{(n)}(c_1^{(n)}), \ldots, g^{(n)}(c_{N_n}^{(n)})$ form a connected partition;

G2: for all i, j $(1 \leq i < j \leq N_n)$, $g^{(n)}(c_i^{(n)}) + g^{(n)}(c_j^{(n)})$ is connected if and only if $c_i^{(n)} + c_j^{(n)}$ is connected.

We remark that, in G2, we have $g^{(n)}(c_i^{(n)}), g^{(n)}(c_j^{(n)}) \in S$ and $c_i^{(n)}, c_j^{(n)} \in A$. Hence, different senses of "+" and "connected" apply in the two cases.

CLAIM 2.107 *For all $n \in \mathbb{N}$, $w^{(n)} \neq \emptyset$.*

Proof For the proof of this claim, we shall drop the n-sub- and superscripts and write N for N_n and c_i for $c_i^{(n)}$. Let G be the neighbourhood graph on $c_1, \ldots, c_N$. By Claim 2.106, G is planar. By Axioms 6 and Claim 2.61, G is connected. Let H be a drawing of G in $\mathbb{S}^2$ (under some mapping $\epsilon : V(G) \to V(H)$); we may assume that H is piecewise linear. By Proposition 2.69, let H^* be a geometric dual of H, which we may likewise assume to be piecewise linear. By Lemma 2.70, every vertex of H lies in exactly one face of H^*. It follows that every edge of H^* is on the boundary of two faces; moreover, H^* by construction contains no isolated nodes. By Theorem 2.82, the faces of H^* form a connected partition in S. So define $g(c_i)$ to be the face of H^* containing the H-vertex $\epsilon(c_i)$. Properties G1 and G2 are then almost immediate. QED

CLAIM 2.108 *Let $I \subseteq \{1, \ldots, N_n\}$, and let $g^{(n)} \in w^{(n)}$. Then $\sum_{i \in I} c_i$ is connected if and only if $\sum_{i \in I} g^{(n)}(c_i)$ is connected.*

Proof Claim 2.61 and property G2. QED

Suppose $n > m \geq 0$, so that $\bar{c}^{(n)}$ refines $\bar{c}^{(m)}$. For all i ($1 \leq i \leq N_m$), let $c_{i,1}, \ldots, c_{i,M_i}$ be the collection of elements of $\bar{c}^{(n)}$ which sum to $c_i^{(m)}$. If $g^{(n)} \in w^{(n)}$, then, we may define the *restriction* of $g^{(n)}$ to $\bar{c}^{(m)}$, written $g^{(n)}|_m$, as follows:

$$g^{(n)}|_m(c_i^{(m)}) = g^{(n)}(c_{i,1}^{(n)}) + \ldots + g^{(n)}(c_{i,M_i}^{(n)})$$

CLAIM 2.109 *Let $g^{(n)} \in w^{(n)}$ with $0 \leq m < n$. Then $g^{(n)}|_m \in w^{(m)}$.*

Proof We must prove that G1 and G2 hold of $g^{(n)}|_m$. G1 is trivial. For G2, we note that, by construction,

$$\begin{aligned} g^{(n)}|_m(c_i^{(m)}) + g^{(n)}|_m(c_j^{(m)}) = g^{(n)}(c_{i,1}^{(n)}) + \cdots + g^{(n)}(c_{i,M_i}^{(n)}) \\ + g^{(n)}(c_{j,1}^{(n)}) + \cdots + g^{(n)}(c_{j,M_j}^{(n)}). \end{aligned}$$

By Claim 2.108, this element of S is connected if and only if the element of A

$$c_{i,1}^{(n)} + \cdots + c_{i,M_i}^{(n)} + c_{j,1}^{(n)} + \cdots + c_{j,M_j}^{(n)} = c_i^{(m)} + c_j^{(m)}$$

is connected. Hence G2 holds as required. QED

CLAIM 2.110 *Let $g \in w^{(n)}$. Then there exists a $g' \in w^{(n+1)}$ such that $g'|_n = g$.*

Proof Choose any $g'' \in w^{(n+1)}$. By Claim 2.109, $g''|_n \in w^{(n)}$. Letting $\bar{r} = g(c_1), \ldots, g(c_{N_n})$ and $\bar{s} = g''|_n(c_1), \ldots, g''|_n(c_{N_n})$, we see that $\bar{r}$ and $\bar{s}$ are c^3-partitions in S with the same neighbourhood graphs—namely, the neighbourhood graph of $c_1, \ldots, c_{N_n}$. By Theorem 2.78, let $h : \mathbb{S}^2 \to \mathbb{S}^2$ be a homeomorphism taking $\bar{s}$ to $\bar{r}$. Thus, $h \circ g''$ maps $\bar{c}^{(n+1)}$ to the faces of a plane graph G in $\mathbb{S}^2$ whose edges include the frontiers of the elements $\bar{r}$. Now let $h' : \mathbb{S}^2 \to \mathbb{S}^2$ be a deformation making all the curved edges of G piecewise linear, while leaving any already piecewise linear edges unaffected. By Theorem 2.82, $g' = h' \circ h \circ g'' \in w^{(n+1)}$ maps $\bar{c}^{(n+1)}$ to an N_{n+1}-tuple in S and it is easy to see that g' satisfies the conditions of the claim. QED

By Claim 2.110, there exists a sequence of embeddings:

$$\emptyset = g^{(0)}, g^{(1)}, g^{(2)}, \ldots$$

such that, for all n $(0 < n)$, $g^{(n)}$ maps $\bar{c}^{(n)}$ to S, and, for all m, n $(0 \leq m < n)$, $g^{(n)}|_m = g^{(m)}$.

Now let $a \in A$ be such that $a = c^{(n)}_{i_1} + \cdots + c^{(n)}_{i_k}$. Then we define

$$g(a) = g^{(n)}(c^{(n)}_{i_1}) + \cdots + g^{(n)}(c^{(n)}_{i_k}) \ .$$

The fact that $g^{(n)}|_m = g^{(m)}$ whenever $0 \leq m < n$ means that this mapping is well defined. It is easy to see that $g : A \to S$ is a Boolean algebra isomorphism; moreover, by Claim 2.108, $g(a)$ is connected if and only if a is connected. That is, we have proved:

CLAIM 2.111 *$\mathfrak{A}$ can be isomorphically embedded in* $\mathrm{ROP}(\mathbb{S}^2)$*, regarded as a $\{c, \leq\}$-structure.*

In view of Claim 2.111, and in order to simplify notation, we might as well take $\mathfrak{A}$ to *be* a substructure of $\mathrm{ROP}(\mathbb{S}^2)$. Note that the previously distinct uses of the Boolean functions and the term "connected" become unambiguous, as do "connected partition", "c^h-partition", "neighbour", and so on. Moreover, since $A \subseteq S$, we may meaningfully talk about the *frontier* $\mathcal{F}(a)$ of any $a \in A$, and apply all the results established previously about elements of $\mathrm{ROP}(\mathbb{S}^2)$. For example, by Lemma 2.73, if $r_1, \ldots, r_n$ is a c^2-partition in A radial about r_1 such that r_1 has at least 2 neighbours, then, for any neighbour r_i of r_1, $\mathcal{F}(r_1) \cap \mathcal{F}(r_i)$ is a Jordan arc. Recall that, for tuples $\bar{r}$ and $\bar{s}$ from $\mathrm{ROP}(\mathbb{S}^2)$, we write $\bar{r} \sim \bar{s}$ if $\bar{r}$ and $\bar{s}$ are similarly situated (in $\mathbb{S}^2$).

Stage 3: In the previous stage, we established that $\mathfrak{A}$ can be chosen to be a substructure of $\mathrm{ROP}(\mathbb{S}^2)$. In this stage, we show that, in that case, $\mathfrak{A}$ is in fact an elementary substructure of $\mathrm{ROP}(\mathbb{S}^2)$.

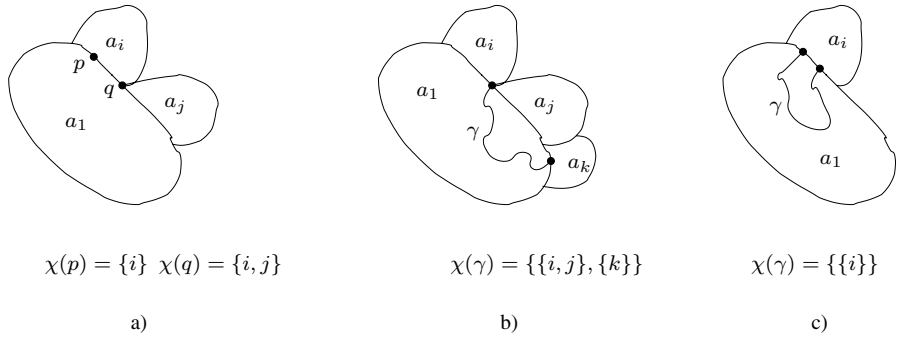


Figure 2.14. The "hub" a_1 of a radial partition.

CLAIM 2.112 *Let $a_1, \ldots, a_n \in A$ be a c^2-partition radial about a_1 such that a_1 has at least 3 neighbours. Let $r_1, r_2 \in S$ be disjoint 2-cells with $a_1 = r_1 + r_2$. Then there exist $c_1, c_2 \in A$ such that $a_1, \ldots, a_n, c_1, c_2 \sim a_1, \ldots, a_n, r_1, r_2$.*

Proof Since a_1, r_1, r_2 are 2-cells with a_1 equal to the disjoint sum of r_1 and r_2, r_1 and r_2 must be separated by a cross-cut γ in a_1. For any neighbour a_i of a_1, $\mathcal{F}(a_1) \cap \mathcal{F}(a_i)$ is a Jordan arc. Let $p \in \mathcal{F}(a_1)$. By inspection, p lies on either one or two Jordan arcs of the form $\mathcal{F}(a_1) \cap \mathcal{F}(a_i)$ where a_i is a neighbour of a_1. We define the *character of* p, written $\chi(p)$ to be the set of those i ($2 \leq i \leq n$) such that a_i is a neighbour of a_1 and $p \in \mathcal{F}(a_i)$ (Fig. 2.14a). Note that $\chi(p)$ has either 1 or 2 elements. If $\chi(p)$ has one element, then p lies on some Jordan arc $\mathcal{F}(a_1) \cap \mathcal{F}(a_i)$, but not at its endpoints. If $\chi(p)$ has two elements, then since a_1 has at least three neighbours, $\chi(p)$ determines p. Now let γ be a cross-cut in a_1. We define the *character of* γ, written $\chi(\gamma)$ to be the set of characters of its endpoints. (See Fig. 2.14b and Fig. 2.14c for examples.) It is routine to show that, if γ_1 and γ_2 are two such cross-cuts and $\chi(\gamma_1) = \chi(\gamma_2)$, there is a homeomorphism of the closed plane onto itself taking a_i to itself for all i ($1 \leq i \leq n$) and taking γ_1 to γ_2. So, to prove the lemma, it suffices to establish that, if γ_1 is any cross-cut in a_1, there exist disjoint 2-cells $c_1, c_2 \in A$ with $a_1 = c_1 + c_2$ such that the cross-cut γ_2 separating c_1 and c_2 in a_1 satisfies $\chi(\gamma_1) = \chi(\gamma_2)$.

Let the endpoints of γ_1 be p and q. We prove the result for the special case where $\chi(\gamma)$, $\chi(p)$ and $\chi(q)$ all contain two elements; the other cases are dealt with similarly. Fig. 2.15a shows the sub-case where $\chi(p)$ and $\chi(q)$ are non-disjoint; Fig. 2.15b shows the sub-case where $\chi(p)$ and $\chi(q)$ are disjoint.

The sub-case of Fig. 2.15a is trivial: Axiom 8 with a_1 substituted for x and a_i for y immediately guarantees the existence of $c_1, c_2 \in A$ partitioning a_1, and

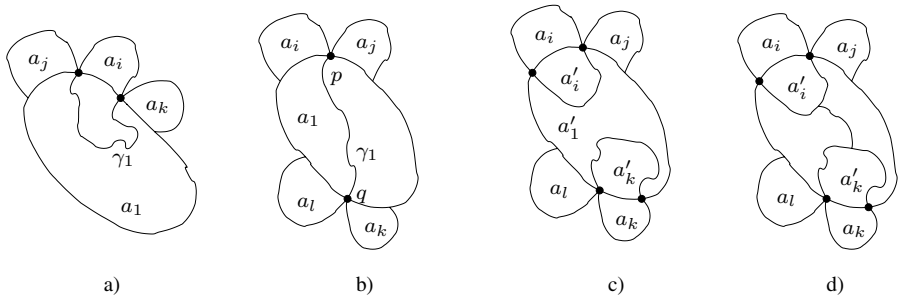


Figure 2.15. The construction of a cross-cut with a given character.

hence separated by a cross-cut γ_2; moreover the connectivity conditions on c_1 and c_2 mean that γ_1 and γ_2 have the same endpoints, so that $\chi(\gamma_1) = \chi(\gamma_2)$.

The sub-case of Fig. 2.15b requires a little more work. However, two applications of Axiom 8 guarantee the existence in A of the regions a'_i, a'_k as in Fig. 2.15c. Axiom 7 then guarantees that the region labelled a'_1 in Fig. 2.15c can be split into two regions as shown in Fig. 2.15d. Summing together appropriate subdivisions of a_1 produces $c_1, c_2 \in A$ separated by an arc γ_2 satisfying $\chi(\gamma_1) = \chi(\gamma_2)$. QED

The rest of this section is devoted to showing that we can relax the conditions of Claim 2.112.

CLAIM 2.113 *Let $n > 1$ and let $a_1, \ldots, a_n \in A$ be a partition such that a_1 is a 2-cell. Let $r_1, r_2 \in S$ be disjoint 2-cells with $a_1 = r_1 + r_2$. Then there exist $c_1, c_2 \in A$ such that $a_1, \ldots, a_n, c_1, c_2 \sim a_1, \ldots, a_n, r_1, r_2$.*

Proof Immediate given claims 2.65 and 2.112. QED

CLAIM 2.114 *Let $n > 1$ and let $a_1, \ldots, a_n \in A$ be a partition such that a_1 is a 2-cell. Let $r \in S$ be such that $r \leq a_1$. Then there exists $c \in A$ such that $a_1, \ldots, a_n, c \sim a_1, \ldots, a_n, r$.*

Proof By the construction of $S = \mathrm{ROP}(\mathbb{S}^2)$, we can partition a_1 into 2-cells $r_1, \ldots, r_m$ such that r can be expressed as the sum of various r_j. It suffices to show that there are $c_1, \ldots, c_m \in A$ such that

$$a_1, \ldots, a_n, r_1, \ldots, r_m \sim a_1, \ldots, a_n, c_1, \ldots, c_m.$$

We proceed by induction on m. If $m = 1$, then $r_1 = a_1$ and we are done. If $m > 1$, by Corollary 2.67, we can renumber the r_i if necessary so that r_1 and

$r_2' = r_2 + \ldots + r_m$ are 2-cells. By Claim 2.113, there exist $c_1, c_2' \in A$ such that $a_1, \ldots, a_n, r_1, r_2' \sim a_1, \ldots, a_n, c_1, c_2'$. Let h be a homeomorphism of the closed plane onto itself mapping a_i to itself, r_1 to c_1 and r_2' to c_2'. By exactly the same argument as for Lemma 2.91, h can be chosen so that $h(r_i) \in S$ for all i ($2 \leq i \leq m$). But then the $h(r_i)$ partition the 2-cell c_2' into 2-cells. So consider the partition $c_2', c_1, a_2, \ldots, a_n$. By inductive hypothesis, there exist $c_2, \ldots, c_m \in A$ such that

$$c_2', c_1, a_2, \ldots, a_n, h(r_2), \ldots, h(r_m) \sim c_2', c_1, a_2, \ldots, a_n, c_2, \ldots, c_m \ .$$

The result then follows. QED

CLAIM 2.115 *Let $n > 1$ and let $a_1, \ldots, a_n \in A$ be a c^2-partition. Let $r \in S$. Then there exists $c \in A$ such that $a_1, \ldots, a_n, c \sim a_1, \ldots, a_n, r$.*

Proof Write r as the sum $r \cdot a_1 + \cdots + r \cdot a_n$ By considering these terms separately, we use Claim 2.114 and an induction similar to that used in the proof of Claim 2.114. The details are routine. QED

CLAIM 2.116 *Let $n \geq 0$ and let $a_1, \ldots, a_n \in A$. Let $r \in S$. Then there exists $c \in A$ such that $a_1, \ldots, a_n, r \sim a_1, \ldots, a_n, c$.*

Proof Immediate given Claims 2.54, 2.63 and 2.115.

CLAIM 2.117 $\mathfrak{A} \preceq \mathrm{ROP}(\mathbb{S}^2)$.

Proof We certainly have $\mathfrak{A} \subseteq \mathrm{ROP}(\mathbb{S}^2)$. Let $n \geq 0$ and let $\phi(x_1, \ldots, x_n)$ be any formula of the form $\exists y \psi(x_1, \ldots, x_n, y)$. Let $a_1, \ldots, a_n \in A$ such that $\mathrm{ROP}(\mathbb{S}^2) \models \phi[a_1, \ldots, a_n]$. Then there exists $r \in S$ such that $\mathrm{ROP}(\mathbb{S}^2) \models \psi[a_1, \ldots, a_n, r]$. By Claim 2.116, there exists $c \in A$ such that $a_1, \ldots, a_n, r \sim a_1, \ldots, a_n, c$. By Lemmas 2.91 and 2.92, $\mathrm{ROP}(\mathbb{S}^2) \models \psi[a_1, \ldots, a_n, c]$. The claim then follows by Proposition 2.95. QED

By Claim 2.117, $\mathfrak{A}$ and $\mathrm{ROP}(\mathbb{S}^2)$ have the same theory. But by construction, $\mathfrak{A} \models \neg\phi$, whence $\mathrm{ROP}(\mathbb{S}^2) \models \neg\phi$, which completes the proof of Lemma 2.105. QED

COROLLARY 2.118 *All splittable, finitely decomposable mereotopologies over $\mathbb{S}^2$ with curve-selection have the same $L_{c,\leq}$-theory, and hence also the same L_C-theory.*

Thus, while Examples 2.17, 2.18 and 2.100 show that there certainly *are* elementarily inequivalent mereotopologies over $\mathbb{R}^2$ and $\mathbb{S}^2$, Corollary 2.118 indicates that there is nothing like the free-for-all one might initially expect. At least

for the signatures $\{c, \leq\}$ and $\{C\}$, the reference to $\mathbf{T}_\Sigma$ as the *standard* first-order mereotopological theory of the closed plane is justified. Corollary 2.174 generalizes this result to apply to any signature of topological primitives.

For reasons of simplicity (which we trust the reader will appreciate) we have provided an axiomatization of well-behaved plane mereotopologies only for the language $L_{c,\leq}$. It should be clear from the foregoing discussion, however, that an analogous result could be obtained for the language L_C, which as we noted, is more expressive over $\mathrm{ROP}(\mathbb{R}^2)$. Such an axiomatization was developed in Schoop, 1999.

Of course, it is one thing to have an axiomatic characterization of the $L_{c,\leq}$-theory of $\mathrm{ROP}(\mathbb{S}^2)$—quite another to determine whether a given $L_{c,\leq}$-sentence is a member of it. The question therefore arises as to the computational characteristics of this problem. Dornheim, 1998 showed (in effect) that this theory is undecidable and hence (since it is a complete theory), not r.e. It follows that the ω-rule (or some equivalent mechanism) is indispensable in this axiomatization. In fact, Schaefer and Štefankovič, 2004 showed (in effect) that the decision problem for $\mathrm{Th}_{c,\leq}\mathrm{ROP}(\mathbb{S}^2)$ is at least as hard as that of second-order arithmetic. Specifically, Schaefer and Štefankovič effectively encode second-order arithmetic in a first-order language with variables ranging over 2-cells in $\mathbb{R}^2$ and primitive predicates expressing the so-called RCC-relations (see Randell et al., 1992, Egenhofer, 1991; but it is easy to see that that theory can in turn be effectively encoded in $\mathrm{Th}_{c,\leq}\mathrm{ROP}(\mathbb{S}^2)$. Schaefer and Štefankovič also consider the complexity of the quantifier-free fragment of their logic, a problem closely related to the well-known problem of recognizing so-called string-graphs (see e.g. Erlich et al., 1976, Kratochvíl, 1988, and show that it is in NEXPTIME. In Schaefer et al., 2003, this bound is improved to NP—a very surprising result.

6. Spatial mereotopology

In this section, we extend the main results of Sec. 4 to the spatial mereotopology $\mathrm{ROP}(\mathbb{R}^3)$. This material is a tidied up version of Pratt and Schoop, 2002.

6.1 Facts about $\mathrm{ROP}(\mathbb{R}^3)$ and $\mathrm{ROP}(\mathbb{S}^3)$

Recall that a 2-*manifold* is a Hausdorff space locally homeomorphic at every point to the open disc $\mathbf{B}^2$, and that a *surface* is a connected 2-manifold.

LEMMA 2.119 *Let X be either $\mathbb{R}^n$ or $\mathbb{S}^n$, and let M be a mereotopology over X having curve-selection. If $r \in M$ with r and $-r$ both connected, then $\mathcal{F}(r)$ is connected.*

Proof Consider the case $X = \mathbb{R}^n$. Let $r \in M$ be connected and non-empty with connected, non-empty complement, and suppose the closed set $\mathcal{F}(r)$ is

not connected. Let d_1 and d_2 be closed sets partitioning $\mathcal{F}(r)$, and let $p \in r$, $q \in -r$. Since r is connected with connected complement, it is easy to see that the conditions of Proposition 2.19 are fulfilled, so that p and q are connected in $\mathbb{R}^n \setminus (d_1 \cup d_2)$. But this is absurd given that $d_1 \cup d_2 = \mathcal{F}(r)$. The case $X = \mathbb{S}^n$ follows easily. QED

LEMMA 2.120 *Let $r \in \mathrm{ROP}(\mathbb{S}^3)$ be such that r and $-r$ are non-empty and connected, and $\mathcal{F}(r)$ is not a surface. Then the graph K^5 can be drawn in $\mathcal{F}(r)$.*

Proof It is easy to see that $\mathcal{F}(r)$ can be finitely triangulated. Call any point where $\mathcal{F}(r)$ is not locally homeomorphic to $\mathbf{B}^2$ a *bad point*; and call any edge of the triangulation all of whose points are bad a *bad edge*. By the properties of triangulations, any bad point either occurs on a bad edge or else is an isolated bad point at a vertex of the triangulation.

If there is a bad edge, then more than two triangles must share this edge, and the embedding of K^5 in $\mathcal{F}(r)$ proceeds as shown in Fig. 2.16a. Assume, then, that there are no bad edges, but that some vertex p of the triangulation is an isolated bad point. Call two triangles with p as a vertex *neighbours* if they share an edge having p as a vertex. Since all edges are good, these triangles can clearly be arranged into disjoint cycles such that each triangle belongs to the same cycle as its two neighbours. Choose one such cycle. By applying a homeomorphism if necessary, we may assume that this triangle-cycle forms a cone with vertex p as shown in Fig. 2.16b. Since there are only finitely many triangles in the triangulation, we can ensure that we choose a triangle-cycle such that the points inside the tip of the cone either all belong to r or all belong to $-r$. Let s be either r or $-r$ depending on which of these possibilities is realized. Note that, since r is non-empty and connected with non-empty, connected complement, so is s.

Let $t \in \mathrm{ROP}(\mathbb{S}^3)$ be a small element representing the tip of the cone, indicated by the light dotted lines in Fig. 2.16b. Removing t from s visibly does not disconnect s, so that $s \cdot -t$ is connected; moreover, t shares some face with $-s$, so that $t + -s = -(s \cdot -t)$ is also connected. Thus, $s \cdot -t$ is non-empty and connected with non-empty connected complement, whence, by Lemma 2.119, $\mathcal{F}(s \cdot -t)$ is connected. Moreover, since p is bad, there must be at least two triangle-cycles with p as vertex; whence $p \in \mathcal{F}(s \cdot -t)$. Thus we may choose a point q on the base rim of t and connect it to p by a Jordan arc α in $\mathcal{F}(s)$ such that the locus of α is disjoint from $\mathcal{F}(t)$ except for its endpoints, as shown in Fig. 2.16c. The embedding of K^5 in $\mathcal{F}(s) = \mathcal{F}(r)$ then proceeds as depicted. QED

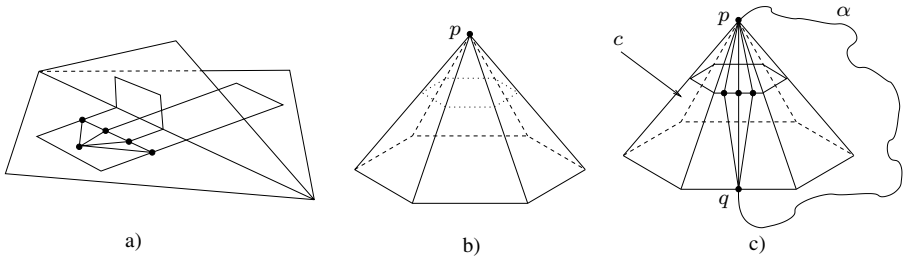


Figure 2.16. Embedding K^5 in non-surfaces (Proof of Lemma 2.120).

One notable difference between $\mathbb{S}^2$ and $\mathbb{S}^3$ is that the Schönflies Theorem, which holds in the former, fails in the latter. In fact, the pathological "region" known as Alexander's horned sphere (Alexander, 1924a) is the best-known counterexample: the frontier of this region is homeomorphic to $\mathbb{S}^2$, but its exterior is not simply connected, and is certainly therefore not homeomorphic to $\mathbf{B}^3$. Nevertheless, Alexander, 1924b also proved a Schönflies-type result for polyhedra, which, in our notation, can be written as follows. (See also Moise, 1977, Ch. 17.)

PROPOSITION 2.121 *Let $r \in \mathrm{ROP}(\mathbb{S}^3)$ be such that $\mathcal{F}(r)$ is homeomorphic to $\mathbb{S}^2$. Then both r^- and $(-r)^-$ are homeomorphic to $\mathbf{D}^3$.*

To avoid cumbersome locutions in the sequel, we define:

DEFINITION 2.122 *Let X be either $\mathbb{R}^n$ or $\mathbb{S}^n$. A* ball *in X is a subset of X similarly situated in X to the unit ball $\mathbf{B}^3$. A* polyhedral ball *in X is a ball which is an element of* $\mathrm{ROP}(X)$.

Thus, if $r \in \mathrm{ROP}(\mathbb{S}^3)$ with $\mathcal{F}(r)$ homeomorphic to $\mathbb{S}^2$, then r and $-r$ are both balls in $\mathbb{S}^3$. Furthermore, if $r \in \mathrm{ROP}(\mathbb{R}^3)$ with $\mathcal{F}(r)$ homeomorphic to $\mathbb{S}^2$ (and hence bounded), then then *exactly one of* r and $-r$ is a ball in $\mathbb{R}^3$. We note in passing:

LEMMA 2.123 *If $r \in \mathrm{RO}(\mathbb{S}^3)$ is a (polyhedral) ball in $\mathbb{S}^3$, then so is $-r$.*

Proof By definition, r is similarly situated in $\mathbb{S}^3$ to $u = \mathbf{B}^3(0,1)$. By considering a spherical inversion, u is similarly situated in $\mathbb{S}^3$ to $-u$. QED

The following well-known theorem will also prove useful in the sequel (see, e.g. Massey, 1967, p. 10).

PROPOSITION 2.124 (CLASSIFICATION THEOREM FOR SURFACES) *Every compact surface is homeomorphic to either (i) $\mathbb{S}^2$ or (ii) the sum of finitely many connected tori or (iii) the sum of finitely many projective planes.*

6.2 Expressing familiar spatial concepts in L_C

Our next task is to show that certain familiar concepts defined on the mereotopology $\mathrm{ROP}(\mathbb{R}^3)$ can be expressed using L_C-formulas. As a preliminary, recall the discussion of Sec. 3.2, which showed that: (i) expressions such as $x^- \cap y^- \cap z \neq \emptyset$ etc. can be regarded as L_C-formulas; and (ii) there is an L_C-formula $\phi_{\mathrm{ci}}(x, y)$ which we may read as "$x^- \cap y^-$ is connected".

Now suppose r and s are elements of $\mathrm{ROP}(\mathbb{R}^3)$, and consider, for example, the set $\mathcal{F}(r) \setminus \mathcal{F}(s)$. Evidently, this set is connected if and only if it is piecewise-linear arc-connected, and therefore if and only if any two points in it are contained within some connected set of the form $r^- \cap t^- \subseteq \mathcal{F}(r) \setminus \mathcal{F}(s)$ with $t \in \mathrm{ROP}(\mathbb{R}^3)$. It follows from the discussion of Sec. 3.2 that there is an L_C-formula satisfied by a pair of regions r, s if and only if $\mathcal{F}(r) \setminus \mathcal{F}(s)$ is connected. In the sequel, then, we write, without further commentary, expressions such as $c(\mathcal{F}(x) \setminus \mathcal{F}(y))$ etc. as L_C-formulas having the obvious interpretations.

LEMMA 2.125 *There exists an L_C-formula $\phi_{K^5}(x)$ such that, for all $r \in \mathrm{ROP}(\mathbb{R}^3)$, $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{K^5}[r]$ if and only if K^5 is embeddable in $\mathcal{F}(r)$.*

Proof The graph K^5 is evidently embeddable in $\mathcal{F}(r)$ if and only if there exist polyhedra v_i ($1 \leq i \leq 5$) and $e_{i,j}$ ($1 \leq i < j \leq 5$), all disjoint from r and from each other, satisfying the following conditions:

1 for all i ($1 \leq i \leq 5$), $v_i^- \cap r^-$ is a singleton;

2 for all i, j ($1 \leq i < j \leq 5$), $e_{i,j}^- \cap r^-$ is connected;

3 for all i, j, i', j' ($1 \leq i < j \leq 5$, $1 \leq i' < j' \leq 5$), $\{i, j\} \cap \{i', j'\} = \emptyset$ implies $e_{i,j}^- \cap e_{i',j'}^- \cap r^- = \emptyset$, and $\{i, j\} \cap \{i', j'\} = \{k\}$ implies $e_{i,j}^- \cap e_{i',j'}^- \cap r^- = v_k^- \cap r^-$.

(Note incidentally that the polyhedra $e_{i,j}$ are not themselves required to be connected—only the sets $e_{i,j}^- \cap r^- = \mathcal{F}(e_{i,j}) \cap \mathcal{F}(r)$.) But the above conditions are expressible in L_C over $\mathrm{ROP}(\mathbb{R}^3)$. QED

LEMMA 2.126 *There exists an L_C-formula $\phi_{\mathrm{b}^*}(x)$ such that, for all $r \in \mathrm{ROP}(\mathbb{R}^3)$:*

1 if $\mathcal{F}(r)$ is connected and unbounded, then $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{b}^*}[r]$;

2 if $\mathcal{F}(r)$ is homeomorphic to $\mathbb{S}^2$, then $\mathrm{ROP}(\mathbb{R}^3) \not\models \phi_{\mathrm{b}^*}[r]$.

Proof Let $\phi_{\mathrm{b}^*}(x)$ be

$$\exists y_1 \exists y_2 (y_1 \cdot x = 0 \wedge y_2 \cdot x = 0 \wedge c(\mathcal{F}(x) \cap \mathcal{F}(y_1) \cap \mathcal{F}(y_2)) \wedge \\ c(\mathcal{F}(x) \setminus \mathcal{F}(y_1)) \wedge c(\mathcal{F}(x) \setminus \mathcal{F}(y_2)) \wedge \neg c(\mathcal{F}(x) \setminus (\mathcal{F}(y_1) \cup \mathcal{F}(y_2)))).$$

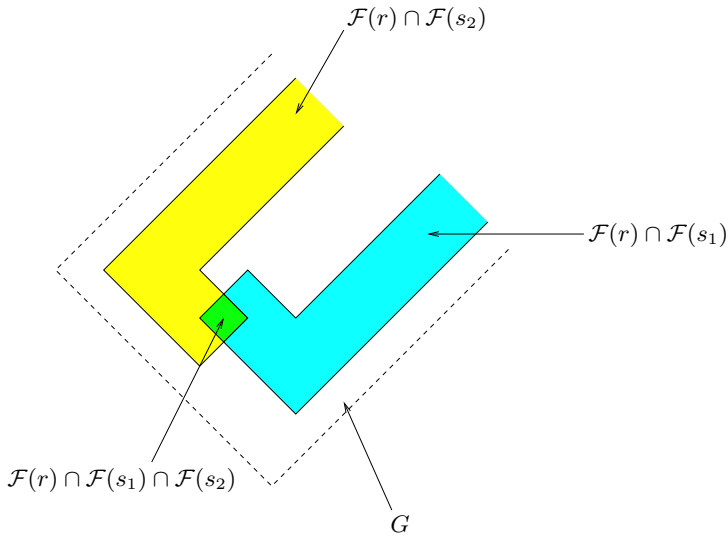


Figure 2.17. Arrangement of $\mathcal{F}(r) \cap \mathcal{F}(s_1)$ and $\mathcal{F}(r) \cap \mathcal{F}(s_2)$ on G (Proof of Lemma 2.126).

Thus, $\phi_{\mathrm{b}^*}(x)$ "says" that there exist polyhedra y_1 and y_2, disjoint from x, such that the sets $\mathcal{F}(x) \cap \mathcal{F}(y_1) \cap \mathcal{F}(y_2)$, $\mathcal{F}(x) \setminus \mathcal{F}(y_1)$ and $\mathcal{F}(x) \setminus \mathcal{F}(y_2)$ are all connected, but the set $\mathcal{F}(x) \setminus (\mathcal{F}(y_1) \cup \mathcal{F}(y_2))$ is not.

Suppose $\mathcal{F}(r)$ is connected and unbounded. Let r be a Boolean combination of finitely many half-spaces, corresponding to a finite set of planes, say, $P_1, \ldots, P_m$; it is then easy to see that $\mathcal{F}(r) \subseteq P_1 \cup \cdots \cup P_m$. Since $\mathcal{F}(r)$ is unbounded, we can draw in $\mathcal{F}(r)$ a rectangular figure G, unbounded on one side (dotted lines in Fig. 2.17), such that G intersects only one of the P_i. Let $s_1, s_2 \in \mathrm{ROP}(\mathbb{R}^3)$ be laminas, infinitely extended in one direction, and placed on G (on the outside of r) so that $\mathcal{F}(r) \cap \mathcal{F}(s_1)$ and $\mathcal{F}(r) \cap \mathcal{F}(s_2)$ are arranged as shown. Since $\mathcal{F}(r)$ is connected, $\mathcal{F}(r) \setminus \mathcal{F}(s_1)$ and $\mathcal{F}(r) \setminus \mathcal{F}(s_2)$ are also connected; and since G lies on just one of the P_i, $\mathcal{F}(r) \setminus (\mathcal{F}(s_1) \cup \mathcal{F}(s_2))$ is not connected. Thus $\mathrm{ROP}(\mathbb{R}^3) \models \gamma[r]$. The second part of the Lemma follows by Proposition 2.47. QED

Let $\phi_{\mathrm{c}}(x)$ be the L_C-formula defined in Lemma 2.27 and satisfied by $r \in \mathrm{ROP}(\mathbb{R}^3)$ if and only if r is connected, and let $\phi_{\mathrm{J}}(x)$ abbreviate the formula $x \neq 0 \wedge x \neq 1 \wedge \phi_{\mathrm{c}}(x) \wedge \phi_{\mathrm{c}}(-x)$.

LEMMA 2.127 *For all $r \in \mathrm{ROP}(\mathbb{R}^3)$, r satisfies $\phi_{\mathrm{J}}(x) \wedge \neg\phi_{K^5}(x) \wedge \neg\phi_{\mathrm{b}^*}(x)$ if and only if $\mathcal{F}(r)$ is homeomorphic to $\mathbb{S}^2$.*

Proof Suppose $\mathcal{F}(r)$ is homeomorphic to $\mathbb{S}^2$. Certainly, by Proposition 2.121, $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{J}}[r]$; by Lemma 2.125, $\mathrm{ROP}(\mathbb{R}^3) \models \neg\phi_{K^5}[r]$; by Lemma 2.126,

$\mathrm{ROP}(\mathbb{R}^3) \models \neg\phi_{\mathrm{b}^*}[r]$. Conversely, suppose that r satisfies $\phi_{\mathrm{J}}(x) \wedge \neg\phi_{K^5}(x) \wedge \neg\phi_{\mathrm{b}^*}(x)$. By Lemma 2.119, $\mathcal{F}(r)$ is connected, and by the first part of Lemma 2.126, $\mathcal{F}(r)$ is bounded. Moreover, K^5 cannot be embedded in $\mathcal{F}(r)$, by Lemma 2.125. Hence $\mathcal{F}(r) = \mathcal{F}(\dot{r})$ is a compact surface, by Lemma 2.120. The result then follows from Proposition 2.124. QED

LEMMA 2.128 *Let $r \in \mathrm{ROP}(\mathbb{R}^3)$ satisfy $\phi_{\mathrm{J}}(x) \wedge \neg\phi_{K^5}(x) \wedge \phi_{\mathrm{b}^*}(x)$. Then r is unbounded.*

Proof Suppose for contradiction that r is bounded, so that we also have $r \in \mathrm{ROP}(\mathbb{S}^3)$. By Lemma 2.120, $\mathcal{F}(r)$ is a surface. Moreover, since r is bounded, $\mathcal{F}(r)$ is compact, and since K^5 cannot be drawn in $\mathcal{F}(s)$, $\mathcal{F}(r)$ is homeomorphic to $\mathbb{S}^2$ by Proposition 2.124. But since $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{b}^*}[r]$, this contradicts the second part of Lemma 2.126. Hence, r is unbounded. QED

LEMMA 2.129 *There exists an L_C-formula $\phi_{\mathrm{b}^3}(x)$ such that, for all $r \in \mathrm{ROP}(\mathbb{R}^3)$, $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{b}^3}[r]$ if and only if r is bounded.*

Proof Let $\phi_{\mathrm{b}^3}(x)$ be the formula

$$\exists y \exists z (x \leq y \wedge y \cdot z = 0 \wedge \\ \phi_{\mathrm{J}}(y) \wedge \neg\phi_{K^5}(y) \wedge \neg\phi_{\mathrm{b}^*}(y) \wedge \phi_{\mathrm{J}}(z) \wedge \neg\phi_{K^5}(z) \wedge \phi_{\mathrm{b}^*}(z)).$$

If r is bounded, let $s \in \mathrm{ROP}(\mathbb{R}^3)$ be a ball in $\mathbb{R}^3$ such that $r \leq s$; and let $t \in \mathrm{ROP}(\mathbb{R}^3)$ be a half-space disjoint from s. By Lemma 2.126, $\mathrm{ROP}(\mathbb{R}^3) \models \neg\phi_{\mathrm{b}^*}[s]$ and $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{b}^*}[t]$. Thus, s and t are suitable witnesses for y and z in $\phi_{\mathrm{b}^3}(x)$, so that $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{b}^3}[r]$.

Conversely, suppose that $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{b}^3}[r]$. Let s and t be witnesses for y and z. By Lemma 2.127, $\mathcal{F}(s)$ is homeomorphic to $\mathbb{S}^2$, whence, by Proposition 2.121, exactly one of s and $-s$ is a ball in $\mathbb{R}^3$. By Lemma 2.128, t is unbounded, and so intersects the complement of every ball in $\mathbb{R}^3$. Therefore $-s$ is not a ball in $\mathbb{R}^3$, so s is. Hence, r is bounded. QED

THEOREM 2.130 *There exists a formula $\phi_{\mathrm{B}}(x)$ such that, for all $r \in \mathrm{ROP}(\mathbb{R}^3)$, $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{B}}[r]$ if and only if r is a polyhedral ball in $\mathbb{R}^3$.*

Proof Let $\phi_{\mathrm{B}}(x)$ be

$$\phi_{\mathrm{J}}(x) \wedge \neg\phi_{K^5}(x) \wedge \neg\phi_{\mathrm{b}^*}(x) \wedge \phi_{\mathrm{b}^3}(x),$$

and apply Lemmas 2.127 and 2.129. QED

Thus, with a little effort, we can define certain familiar topological notions over $\mathrm{ROP}(\mathbb{R}^3)$ using L_C-formulas. The following technical material, which is devoted to defining some decidedly *unfamiliar* topological notions over $\mathrm{ROP}(\mathbb{R}^3)$, will be used in the sequel. We recall the discussion of compactifications in Sec. 3.3, and consider the mapping $r \mapsto \dot{r}$ from $\mathrm{ROP}(\mathbb{R}^3)$ to its 1-point compactification $\mathrm{ROP}(\mathbb{S}^3)$. By Lemmas 2.36 and 2.37, this mapping is a Boolean algebra isomorphism and preserves the properties of connectedness and non-connectedness. For technical reasons, we will occasionally need to consider properties of elements in $\mathrm{ROP}(\mathbb{R}^3)$ whose defining conditions make reference to their counterparts in $\mathrm{ROP}(\mathbb{S}^3)$.

For all $r \in \mathrm{ROP}(\mathbb{R}^3)$, $\infty \in \dot{r}$ if and only if $-r$ is bounded, and $\infty \in \dot{r}^-$ if and only if r is unbounded (where the closure operator $^-$ refers to the topology on $\mathbb{S}^3$). By Lemma 2.129 then, it is harmless to employ the expression $\infty \in \dot{x}$ in L_C-formulas, since we can take it as a mnemonic for $\phi_{\mathrm{b}^3}(-x)$; and similarly for expressions such as $\infty \in \dot{x}^-$, $\infty \in \mathcal{F}(\dot{x})$, etc.

LEMMA 2.131 *There exists a formula $\phi_{\dot{K^5}}(x)$ satisfied by $r \in \mathrm{ROP}(\mathbb{R}^3)$ if and only if K^5 is embeddable in $\mathcal{F}(\dot{r})$.*

Proof As for Lemma 2.125, making the obvious adjustments to accommodate the point at infinity. QED

LEMMA 2.132 *There exists a formula $\phi_{\dot{\mathrm{B}}}(x)$ such that, for all $r \in \mathrm{ROP}(\mathbb{R}^3)$, $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\dot{\mathrm{B}}}[r]$ if and only if $\dot{r}$ is a ball in $\mathbb{S}^3$.*

Proof Let $\phi_{\dot{\mathrm{B}}}(x)$ be $\phi_{\mathrm{J}}(x) \wedge \neg\phi_{\dot{K^5}}(x)$. If $\dot{r}$ is a ball in $\mathbb{S}^3$, it is evident that $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\dot{\mathrm{B}}}[r]$. Conversely, suppose $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\dot{\mathrm{B}}}[r]$. By Lemmas 2.120 and 2.131, $\mathcal{F}(\dot{r})$ is a surface in $\mathbb{S}^3$. Furthermore, by Proposition 2.124, $\mathcal{F}(\dot{r})$ is homeomorphic to $\mathbb{S}^2$. The result then follows by Proposition 2.121. QED

6.3 Characterizing triangulations in L_C

In Sec. 4, we showed that every tuple in $\mathrm{ROP}(\mathbb{R}^2)$ satisfies a topologically complete L_C-formula—that is, an L_C-formula with the property that all tuples satisfying it are similarly situated. Our proof exploited Whitney's theorem on 3-connected graphs in the plane to show that any c^3-partition in $\mathrm{ROP}(\mathbb{S}^2)$ is determined up to similar situation by its neighbourhood graph. However, Whitney's theorem is not available for $\mathbb{S}^3$, and so we must adopt an alternative approach to analysing the expressive power of L_C over $\mathrm{ROP}(\mathbb{R}^3)$. This approach has the advantage of being, in some ways, more straightforward than that of Sec. 4, though the topologically complete formulas it constructs are more complicated.

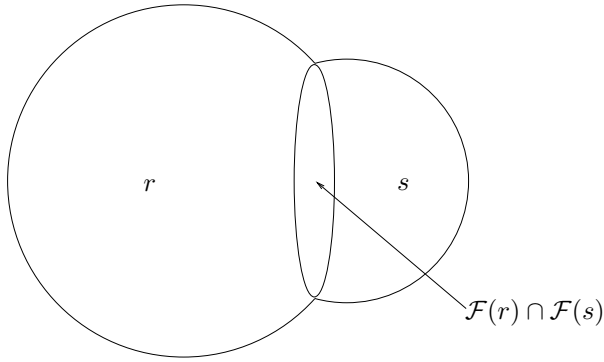


Figure 2.18. The configuration of Proposition 2.133.

We assume familiarity with the basic theory of triangulations and PL-complexes: for details, see, e.g., Moise, 1977, Ch. 7. We also require the following "obvious" result about balls in $\mathbb{S}^3$ (Pratt and Schoop, 2002, Theorem 3.14).

PROPOSITION 2.133 *Let $r, s \in \mathrm{RO}(\mathbb{S}^3)$ be disjoint balls in $\mathbb{S}^3$ such that $r+s$ is also a ball in $\mathbb{S}^3$. Then $\mathcal{F}(r) \cap \mathcal{F}(s) \cap \mathcal{F}(r+s)$ is the locus of a Jordan curve, and $\mathcal{F}(r) \cap \mathcal{F}(s)$ is homeomorphic to the closed disc $\mathbf{D}^2$.*

The situation is illustrated in Fig. 2.18.

DEFINITION 2.134 *A quadruple $q = \langle r_1, r_2, r_3, r_4 \rangle$ of pairwise disjoint elements of $\mathrm{ROP}(\mathbb{S}^3)$ is a* q-cell in $\mathbb{S}^3$ *if, for all non-empty $J \subseteq \{1, 2, 3, 4\}$, the polyhedron $\sum_{j \in J} r_j$ is a ball in $\mathbb{S}^3$.*

The reference to the containing space $\mathbb{S}^3$ is significant: in the sequel, we introduce q-cells in $\mathbb{R}^3$. However, we sometimes speak simply of *q-cells* if it is clear from context which space we are talking about (or if it makes no difference).

EXAMPLE 2.135 *Consider the regular open tetrahedron t_0 with vertices $v_1 = (0,0,0)$, $v_2 = (1,0,0)$, $v_3 = (0,1,0)$, $v_4 = (0,0,1)$. Let t_1, t_2, t_3, t_4 be the four regular open tetrahedra (taken in some fixed order) each having three vertices from $\{v_1, \ldots, v_4\}$ and the point $(1/4, 1/4, 1/4)$ as the fourth vertex (Fig. 2.19). Evidently, the quadruple $q_0 = \langle t_1, t_2, t_3, t_4 \rangle$ is a q-cell.*

THEOREM 2.136 *All q-cells in $\mathbb{S}^3$ are similarly situated in $\mathbb{S}^3$.*

Proof Let $\langle a, b, c, d \rangle$ be a q-cell. Since a, b, c, $a+b$, $b+c$, $a+c$ and $a+b+c$ are balls, by Proposition 2.133, the sets $\mathcal{F}(a) \cap \mathcal{F}(b)$, $\mathcal{F}(a) \cap \mathcal{F}(c)$, $\mathcal{F}(b) \cap \mathcal{F}(c)$ and $\mathcal{F}(a+b) \cap \mathcal{F}(c)$ are all closed discs. Letting $S = \mathcal{F}(a+b+c)$, it is then

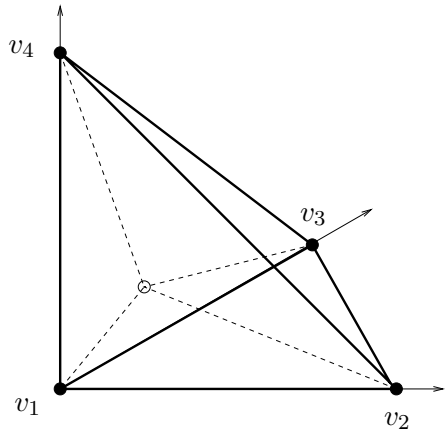


Figure 2.19. The q-cell q_0.

easy to show that the sets $\mathcal{F}(a) \cap S$, $\mathcal{F}(b) \cap S$ and $\mathcal{F}(c) \cap S$ must be arranged on S as shown in Fig. 2.20a, up to similar situation. Let $e = -(a+b+c+d)$; then, by Lemma 2.123, $\sum(B \cup \{e\})$ is a ball for any proper subset $B \subset \{a, b, c, d\}$. Thus, all of the sets $a+b+c$, d, e, $a+b+c+d$ and $a+b+c+e$ are balls. By Proposition 2.133 again, $\mathcal{F}(d) \cap S$ and $\mathcal{F}(e) \cap S$ are both closed discs, whose common frontier in the space S is the locus of some Jordan curve γ, say.

Consider how γ might be drawn on S. Since $a + d$ and $a + e$ are balls, by Proposition 2.133, $\mathcal{F}(a) \cap \mathcal{F}(d)$ and $\mathcal{F}(a) \cap \mathcal{F}(e)$ are closed discs. Similarly, $\mathcal{F}(b) \cap \mathcal{F}(d)$, $\mathcal{F}(b) \cap \mathcal{F}(e)$, $\mathcal{F}(c) \cap \mathcal{F}(d)$ and $\mathcal{F}(c) \cap \mathcal{F}(e)$ are closed discs. Hence γ divides each of the three sets $\mathcal{F}(a) \cap S$, $\mathcal{F}(b) \cap S$ and $\mathcal{F}(c) \cap S$ into two residual domains. Moreover, γ cannot pass through either of the points X or Y in Fig. 2.20a; for otherwise, one of the sets $\mathcal{F}(a) \cap \mathcal{F}(d)$, $\mathcal{F}(b) \cap \mathcal{F}(d)$, $\mathcal{F}(c) \cap \mathcal{F}(d)$, $\mathcal{F}(a) \cap \mathcal{F}(e)$, $\mathcal{F}(b) \cap \mathcal{F}(e)$ or $\mathcal{F}(c) \cap \mathcal{F}(e)$ would fail to be a closed disc. It is then easy to see that γ and the region $\mathcal{F}(d) \cap S$ it encloses must lie in S as shown in Fig. 2.20b or Fig. 2.20c, up to similar situation. But these two arrangements of a, b, c, d are mirror images. QED

NOTATION 2.137 *If $q = \langle t_1, \ldots, t_4 \rangle$ is a q-cell, denote the component polyhedron t_i by $q[i]$ for all i $(1 \leq i \leq 4)$. Denote the polyhedron $t_1 + \cdots + t_4$ by $\hat{q}$.*

In Example 2.135, $\hat{q}_0$ is the interior of the convex hull of the points $V = \{v_1, \ldots, v_4\}$. We employ familiar terms from discussions of simplicial complexes: a *face* of q_0 is the convex closure of any non-empty subset of V; a face of q_0 is proper if it is not the whole of $(\hat{q}_0)^-$; a *vertex* of q_0 is an element of V.

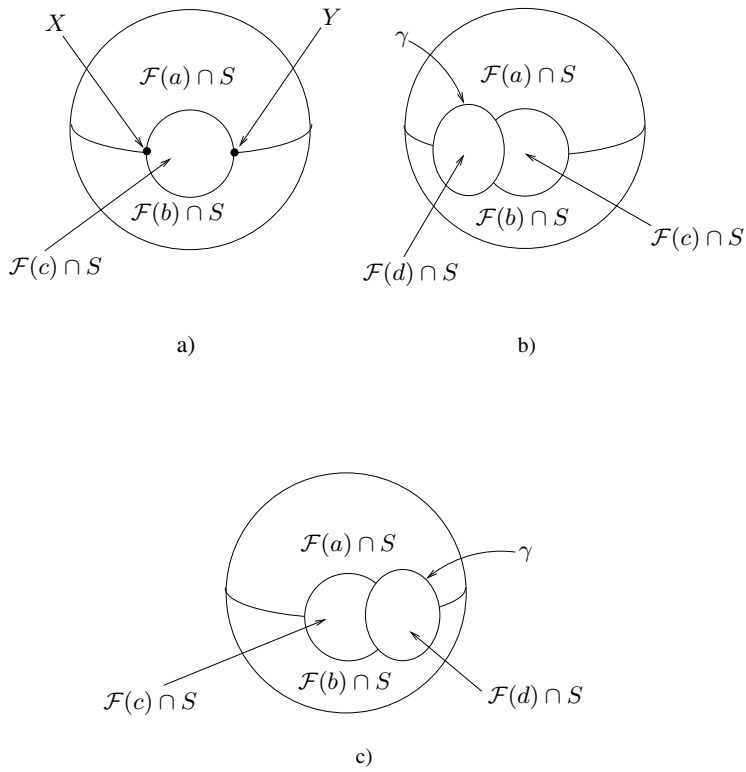


Figure 2.20. Possible arrangements of $\mathcal{F}(a) \cap S$, $\mathcal{F}(b) \cap S$, $\mathcal{F}(c) \cap S$ and $\mathcal{F}(d) \cap S$, where $S = \mathcal{F}(a + b + c)$ (Proof of Theorem 2.136).

DEFINITION 2.138 *Let q be any q-cell, and h a homeomorphism of $\mathbb{S}^3$ onto itself taking q_0 to q. A* (proper) face *of q is a set of points $h(F)$, where F is a* (*proper*) *face of q_0. A* vertex *of q is a point $h(v)$, where v is a vertex of q_0.*

We remark that, in Definition 2.138, a suitable homeomorphism h can always be found, by Theorem 2.136; moreover, since the faces of q_0 are expressible as set-algebraic combinations of the polyhedra $t_1, \ldots, t_4$ and their topological closures, the precise choice of h does not matter. Thus, q-cells are simply homeomorphic images of the q-cell q_0 of Example 2.135, with the notions of *face* and *vertex* transferred in the obvious way.

DEFINITION 2.139 *A* q-cell partition (*in* $\mathrm{ROP}(\mathbb{S}^3)$) *is a sequence $\bar{q} = q_1, \ldots, q_n$ of q-cells in $\mathbb{S}^3$ such that* (*i*) *$\hat{q}_1, \ldots, \hat{q}_n$ is a partition in* $\mathrm{ROP}(\mathbb{S}^3)$*; and* (*ii*) *for all i, j $(1 \leq i < j \leq n)$, if F is a face of q_i and G a face of q_j, then $F \cap G$ is either empty or a face of both q_i and q_j. A* vertex *of a q-cell partition is a vertex of one of its elements.*

Thus, q-cell partitions define (finite) PL-complexes in the obvious way: each q-cell in the partition corresponds to a PL 3-simplex, and its proper faces to PL d-simplices for $d < 3$.

DEFINITION 2.140 *Let $\bar{q} = q_1, \ldots, q_N$ and $\bar{q}' = q'_1, \ldots, q'_N$ be q-cell partitions in* $\mathrm{ROP}(\mathbb{S}^3)$*. We say that $\bar{q}$ and $\bar{q}'$ are* isomorphic *if there is a bijection between the vertices of $\bar{q}$ and the vertices of $\bar{q}'$ such that, for all i, j $(1 \leq i \leq N$, $1 \leq j \leq 4)$, the vertices of q_i lying on the frontier of $q_i[j]$ are mapped to the vertices of q'_i lying on the frontier of $q'_i[j]$.*

LEMMA 2.141 *Isomorphic q-cell partitions in* $\mathrm{ROP}(\mathbb{S}^3)$ *are similarly situated in $\mathbb{S}^3$.*

Proof Isomorphic q-cell partitions define isomorphic PL-complexes. QED

We conclude this subsection by extending the notions of q-cell and q-cell partition to the open space $\mathbb{R}^3$.

DEFINITION 2.142 *A quadruple $q = \langle r_1, r_2, r_3, r_4 \rangle$ of elements of* $\mathrm{ROP}(\mathbb{R}^3)$ *is a* q-cell in $\mathbb{R}^3$ *if $\dot{q} = \langle \dot{r}_1, \dot{r}_2, \dot{r}_3, \dot{r}_4 \rangle$ is a q-cell in $\mathbb{S}^3$. A sequence $\bar{q} = q_1, \ldots, q_n$ of q-cells in $\mathbb{R}^3$ is a* q-cell-partition in $\mathrm{ROP}(\mathbb{R}^3)$ *if $\dot{q}_1, \ldots, \dot{q}_n$ is a q-cell partition in* $\mathrm{ROP}(\mathbb{S}^3)$.

DEFINITION 2.143 *Let $\bar{q} = q_1, \ldots, q_n$ and $\bar{q}' = q'_1, \ldots, q'_n$ be q-cell partitions in* $\mathrm{ROP}(\mathbb{R}^3)$*. We say that $\bar{q}$ and $\bar{q}'$ are* isomorphic *if:* (*i*) *the corresponding q-cell partitions $\dot{q}_1, \ldots, \dot{q}_n$ and $\dot{q}'_1, \ldots, \dot{q}'_n$ in* $\mathrm{ROP}(\mathbb{S}^3)$ *are isomorphic; and* (*ii*) *for all i, j $(1 \leq i \leq n$, $1 \leq j \leq 4)$, $q_i[j]$ is bounded if and only if $q'_i[j]$ is bounded.*

Intuitively, knowing which $q_i[j]$ are bounded for a q-cell partition $q_1, \ldots, q_n$ in $\mathrm{ROP}(\mathbb{R}^3)$ amounts to knowing, up to homeomorphism, where the point at infinity is in the corresponding q-cell partition in $\mathrm{ROP}(\mathbb{S}^3)$. More precisely, we have:

LEMMA 2.144 *Let $\bar{q} = q_1, \ldots q_n$ and $\bar{q}' = q'_1, \ldots, q'_n$ be similarly situated q-cell partitions in $\mathrm{ROP}(\mathbb{S}^3)$. Let $p \in \mathbb{S}^3$ such that, for all i, j $(1 \leq i \leq n, 1 \leq j \leq 4)$, $p \in (q_i[j])^-$ if and only if $p \in (q'_i[j])^-$. Then there is a homeomorphism $h : \mathbb{S}^3 \to \mathbb{S}^3$ fixing p and mapping $\bar{q}$ to $\bar{q}'$.*

Proof Parallel to the proof of Lemma 2.88. QED

THEOREM 2.145 *Isomorphic q-cell partitions in $\mathrm{ROP}(\mathbb{R}^3)$ are similarly situated in $\mathbb{R}^3$.*

Proof Let $q_1, \ldots, q_n$ and $q'_1, \ldots, q'_n$ be isomorphic q-cell partitions in $\mathrm{ROP}(\mathbb{R}^3)$. Then $\dot{q}_1, \ldots, \dot{q}_n$ and $\dot{q'}_1, \ldots, \dot{q'}_n$ are isomorphic q-cell partitions such that, for all i, j $(1 \leq i \leq n, 1 \leq j \leq 4)$, $\infty \in (\dot{q}_i[j])^-$ if and only if $\infty \in (\dot{q'}_i[j])^-$. By Lemmas 2.141 and 2.144, there exists a homeomorphism h of $\mathbb{S}^3$ onto itself mapping $\dot{q}_1, \ldots, \dot{q}_n$ to $\dot{q'}_1, \ldots, \dot{q'}_n$, and fixing ∞. Thus, $h' = h \setminus \{\langle \infty, \infty \rangle\}$ is a homeomorphism of $\mathbb{R}^3$ onto itself mapping $q_1, \ldots, q_n$ to $q'_1, \ldots, q'_n$. QED

6.4 Expressive power of L_C in $\mathrm{ROP}(\mathbb{R}^3)$

We are now ready to show that every tuple in $\mathrm{ROP}(\mathbb{R}^3)$ satisfies a formula which is topologically complete in $\mathrm{ROP}(\mathbb{R}^3)$ over $\mathbb{R}^3$.

LEMMA 2.146 *For all $N > 0$, there exists a formula $\phi^N_{\mathrm{q}}(\bar{z})$ such that, for any $4N$-tuple $\bar{t}$ from $\mathrm{ROP}(\mathbb{R}^3)$, $\mathrm{ROP}(\mathbb{R}^3) \models \phi^N_{\mathrm{q}}[\bar{t}]$ if and only if $\bar{t}$ is a q-cell partition in $\mathrm{ROP}(\mathbb{R}^3)$.*

Proof Let $\phi_{\dot{\mathrm{B}}}(x)$ be as defined in Lemma 2.132, and suppose $s_1, \ldots, s_4 \in \mathrm{ROP}(\mathbb{R}^3)$. Then the quadruple $\langle \dot{s}_1, \ldots, \dot{s}_4 \rangle$ is a q-cell in $\mathbb{S}^3$ if and only if $\mathrm{ROP}(\mathbb{R}^3) \models \phi_{\mathrm{q}}[s_1, \ldots, s_4]$, where $\phi_{\mathrm{q}}(y_1, \ldots, y_4)$ is the formula

$$\bigwedge \Big\{ \phi_{\dot{\mathrm{B}}}\Big(\sum_{j \in J} y_j \Big) \mid \emptyset \neq J \subseteq \{1, 2, 3, 4\} \Big\}.$$

The result then follows easily. QED

LEMMA 2.147 *Let $\bar{t}$ be a $4N$-tuple forming an N-element q-cell partition in $\mathrm{ROP}(\mathbb{R}^3)$. Then we can find a formula $\gamma(\bar{z})$ such that, for any $4N$-tuple $\bar{t}'$ of $\mathrm{ROP}(\mathbb{R}^3)$, $\mathrm{ROP}(\mathbb{R}^3) \models \gamma[\bar{t}']$ if and only if $\bar{t}'$ is an N-element q-cell partition isomorphic to $\bar{t}$.*

Proof Almost immediate from Lemmas 2.129 and 2.146 and the discussion of Sec. 3.2. QED

LEMMA 2.148 *Every q-cell partition in* $\mathrm{ROP}(\mathbb{R}^3)$ *satisfies an* L_C*-formula which is topologically complete in* $\mathrm{ROP}(\mathbb{R}^3)$ *over* $\mathbb{R}^3$.

Proof Theorem 2.145 and Lemma 2.147. QED

LEMMA 2.149 *Any* n*-tuple* $\bar{r}$ *from* $\mathrm{ROP}(\mathbb{R}^3)$ *can be refined to an* N*-element q-cell partition. That is: there exists a* $4N$*-tuple* $\bar{t}$ *from* $\mathrm{ROP}(\mathbb{R}^3)$ *and a* $(4N \times n)$ *Boolean array* A *such that* $\bar{t}$ *forms a q-cell partition in* $\mathrm{ROP}(\mathbb{R}^3)$ *and* $\bar{r} = \bar{t}A$.

Proof By the definition of $\mathrm{ROP}(\mathbb{R}^3)$, we can certainly refine $\bar{r}$ to a partition of convex regions of $\mathbb{R}^3$, each of which is bounded by a finite number of planes, and thence, by triangulating these convex regions, into a partition of polyhedra $t_1, \ldots, t_N$, such that each t_i is a ball in $\mathbb{S}^3$, and the boundary of each t_i $(1 \leq i \leq N)$ is composed of 4 "triangles". By subdividing each t_i, we can construct a q-cell q_i whose faces are exactly the triangles bounding t_i, and such that $\hat{q}_i = t_i$. Then $q_1, \ldots, q_N$ is the required q-cell partition. QED

THEOREM 2.150 *Every tuple in* $\mathrm{ROP}(\mathbb{R}^3)$ *satisfies some* L_C*-formula which is topologically complete in* $\mathrm{ROP}(\mathbb{R}^3)$ *over* $\mathbb{R}^3$.

Proof Let $\bar{r}$ be a tuple from $\mathrm{ROP}(\mathbb{R}^3)$. Let $\bar{t}$ and A be as in Lemma 2.149, and by Lemma 2.148 let $\phi_{\bar{t}}(\bar{z})$ be a formula satisfied by $\bar{t}$ which is topologically complete in $\mathrm{ROP}(\mathbb{R}^3)$ over $\mathbb{R}^3$. Then the formula $\exists \bar{z}(\phi_{\bar{t}}(\bar{z}) \wedge \bar{x} = \bar{z}A)$, which is also topologically complete in $\mathrm{ROP}(\mathbb{R}^3)$ over $\mathbb{R}^3$, is satisfied by $\bar{r}$. QED

This concludes the main business of this section: the language L_C is sufficiently expressive that every tuple of polyhedra in $\mathbb{R}^3$ can be characterized up to the relation of similar situation in $\mathbb{R}^3$ by one of its formulas. Moreover, it is easy to see that an analogous result must obtain for polyhedra in $\mathbb{S}^3$. Of course, the characterizing formulas for tuples of polyhedra obtained in this section are much more complicated than the corresponding $L_{c,\leq}$-formulas for tuples of polygons obtained in Sec. 4.

In Sec. 5, we exploited the high expressive power of $L_{c,\leq}$ in $\mathrm{ROP}(\mathbb{S}^2)$ to obtain an axiomatization of $\mathrm{Th}_{c,\leq}(\mathrm{ROP}(\mathbb{S}^2))$, and thence, a formulation of the conditions under which an arbitrary mereotopology over $\mathbb{S}^2$ has the same $L_{c,\leq}$-theory as $\mathrm{ROP}(\mathbb{S}^2)$. The question therefore arises as to whether an analogous

approach is possible for characterizing "reasonable" spatial mereotopologies using the results of this section. The major disincentive to such an undertaking is the relative weakness of the requirement of finite decomposability in $\mathbb{S}^3$. For the plane case, the requirement of finite decomposability led very easily to the existence of c^3-partition refinements, which paved the way for an axiomatic characterization of $\mathrm{Th}_{c,\leq}(\mathrm{ROP}(\mathbb{S}^2))$. In the spatial case, by contrast, much stronger assumptions are needed to guarantee the existence of q-cell partitions. Thus, while the identification of a standard theory of spatial mereotopology is certainly conceivable, it is not obvious, at the time of writing, how best to approach this matter.

7. Model theory

In Sec. 2, we defined a *mereotopology* over a topological space X to be a Boolean sub-algebra M of $RO(X)$ in which, for all $p \in o \subseteq X$, with o open, there exists $r \in M$ such that $p \in r \subseteq o$. However, we also promised a purely intrinsic characterization of such structures—one making no reference to points or topological spaces. In this section, we fulfil that promise, and (partially) realize the vision with which we started this chapter, of a reconstruction of topology where the fundamental objects are not points, but regions.

7.1 Abstract models of mereotopological theories

We begin by noting some simple facts about mereotopologies over topological spaces of various kinds.

LEMMA 2.151 *Let M be a mereotopology over a topological space X, considered as a structure interpreting the signature $\{C, +, \cdot, -, 0, 1, \leq\}$. (i) The sentences Φ_{CA} consisting of the usual axioms of Boolean algebra together with*

$$\begin{aligned}
&\forall x \neg C(x, 0)\\
&\forall x (x > 0 \rightarrow C(x, x))\\
&\forall x \forall y (C(x, y) \rightarrow C(y, x))\\
&\forall x \forall y (C(x, y) \wedge y \leq z \rightarrow C(x, z))\\
&\forall x \forall y (C(x, y + z) \rightarrow C(x, y) \vee C(x, z))
\end{aligned}$$

are all true in M. (ii) If X is weakly regular, then the sentence ϕ_{ext} given by

$$\forall x \forall y (\forall z (C(x, z) \rightarrow C(y, z)) \rightarrow x \leq y)$$

is true in M. (iii) If X is compact and Hausdorff, then the sentence ϕ_{int} given by

$$\forall x \forall y (\neg C(x, y) \rightarrow \exists z (\neg C(x, -z) \wedge \neg C(y, z)))$$

is true in M.

Proof (i) Straightforward. (ii) Lemma 2.22. (iii) Suppose $r, s \in M$ with $r^- \cap s^- = \emptyset$. Since X is regular, by Lemma 2.23, for each point in $p \in r^-$, there is $r_p \in M$ with $p \in r_p$ and $s^- \subseteq -r_p$. Since the r_p cover r^-, choose a finite subcover, and let the sum of this subcover be t. Then $r^- \subseteq t$ and $s^- \subseteq -t$. QED

The three claims in Lemma 2.151 all have converses. Specifically:

PROPOSITION 2.152 *Let $\mathfrak{A}$ be a structure interpreting the signature $\Sigma = \{C, +, \cdot, -, 0, 1 \leq\}$. (i) If $\mathfrak{A} \models \Phi_{\mathrm{CA}}$, then $\mathfrak{A}$ is isomorphic (as a Σ-structure) to a mereotopology over a topological space X; in fact, X can always be chosen so as to be semi-regular and T_0. (ii) If $\mathfrak{A} \models \Phi_{\mathrm{CA}} \cup \{\phi_{\mathrm{ext}}\}$, then X can be chosen so as to be weakly regular and T_1. (iii) If $\mathfrak{A} \models \Phi_{\mathrm{CA}} \cup \{\phi_{\mathrm{ext}}, \phi_{\mathrm{int}}\}$, then X can be chosen so as to be compact and Hausdorff.*

These results first appeared (in equivalent form) in Dimov and Vakarelov, 2006, Düntsch and Winter, 2005 and Roeper, 1997, respectively. In the literature, structures satisfying Φ_{CA} are sometimes referred to as *contact algebras*, the sentence ϕ_{ext} as the *extensionality axiom*, and the sentence ϕ_{int} as the *interpolation axiom*. Together, Lemma 2.151 and Proposition 2.152 show that mereotopologies over certain classes of topological spaces can be characterized purely intrinsically, without reference to those spaces or the points that make them up. We note in passing that Proposition 2.152 speaks of *mereotopologies over X* (Definition 2.5), while the sources cited refer only to *dense sub-algebras of* $\mathrm{RO}(X)$. This slight strengthening is immediate from the relevant proofs, and improves the match between Lemma 2.151 and Proposition 2.152. For a fuller discussion, see Ch. 3.

Furthermore, it turns out that the topological realizations in Proposition 2.152 (iii) are, in an important sense, unique. We motivate this result with a simple observation.

LEMMA 2.153 *Let M_i be a mereotopology over the topological space X_i, for $i = 1, 2$. Suppose there is a homeomorphism $h : X_1 \to X_2$ which maps M_1 onto M_2. Then, for any signature Σ of topological primitives, h induces a structure isomorphism $h : M_1 \simeq_\Sigma M_2$.*

Proof Immediate. QED

The uniqueness of the topological realizations in Proposition 2.152 (iii) takes the form of a partial converse of Lemma 2.153:

THEOREM 2.154 (ROEPER, 1997) *Let M_i be a mereotopology over a compact, Hausdorff space X_i $(i = 1, 2)$. Suppose there is a structure isomorphism $f : M_1 \simeq_C M_2$. Then there exists a homeomorphism $h : X_1 \to X_2$ which induces f—that is, one such that, for all $r \in M_1$, $f(r) = h(r)$.*

Thus, every model of $\Phi_{\mathrm{CA}} \cup \{\phi_{\mathrm{ext}}, \phi_{\mathrm{int}}\}$ is isomorphic to exactly one mereotopology over a compact, Hausdorff space (up to homeomorphism). Since this fact is important for the development here, we present details of the proof.

We assume familiarity with the theory of ultrafilters: for details, see Koppelberg, 1989, Ch. 1, Sec. 2. In this context, recall that, for B a Boolean algebra, a *filter* on B is a set $F \subseteq B$ such that $a, b \in F$ implies $a \cdot b \in F$, and $a \in F$, $a \leq b \in B$ implies $b \in F$. A filter is *proper* if it is not the whole of B, or equivalently, if it does not contain 0. A proper filter U is an *ultrafilter* if it is maximal under set-inclusion, or equivalently, if $b_1 + b_2 \in U$ implies $b_1 \in U$ or $b_2 \in U$. The following result is standard (Koppelberg, 1989, Ch. 1, 2.16).

PROPOSITION 2.155 (PRIME IDEAL THEOREM) *Any proper filter on a Boolean algebra can be extended to an ultrafilter.*

In the following lemmas, let M be a mereotopology over a compact, Hausdorff space X. Since a compact, Hausdorff space is normal (and hence regular), Lemma 2.23 applies.

LEMMA 2.156 *Let U be an ultrafilter on M. Then the set $\bigcap\{r^- | r \in U\}$ is a singleton. We denote the member of this set by p_U and say that U* converges to *p_U.*

Proof We first show that $\bigcap\{u^- | u \in U\}$ contains at least one point. For otherwise, $\bigcup\{X \setminus u^- | u \in U\} = X$, whence $\{-u | u \in U\}$ covers X. By compactness of X, let $-u_1, \ldots, -u_n$ be a finite subcover. Then $-u_1 + \cdots + -u_n = 1$; i.e. $u_1 \cdot \cdots \cdot u_n = 0 \in U$, contradicting the fact that U is proper. Next we show that $\bigcap\{u^- | u \in U\}$ contains at most one point. For suppose p, q are distinct points of X. By Lemma 2.23, there exists $r \in M$ such that $p \in r$ and $q \in -r$. Hence $p \notin (-r)^-$ and $q \notin r^-$. Since U is an ultrafilter, either r or $-r$ is in U, so that either p or q is not in $\bigcap\{u^- | u \in U\}$. QED

LEMMA 2.157 *Let U be an ultrafilter on M, and let $r \in M$. If $p_U \in r$, then there exists $s \in U$ such that $p_U \in s$ and $s^- \subseteq r$. Hence also, $r \in U$.*

Proof Suppose $p_U \in r \in M$. Then $p_U \notin (-r)^-$, and by Lemma 2.23, there exists $s \in M$ such that $p_U \in s$ and $s^- \subseteq r$. But since $p_U \notin (-s)^-$ we have $-s \notin U$, and thus $s \in U$. QED

DEFINITION 2.158 *If U and V are ultrafilters on M, we say U and V are* contacting *if $r^- \cap s^- \neq \emptyset$ for all $r \in U$ and $s \in V$.*

LEMMA 2.159 *If U and V are ultrafilters on M, then U and V are contacting if and only if $p_U = p_V$.*

Proof The if-direction is trivial. For the only-if direction, suppose that $p_U \neq p_V$. By Lemma 2.23, there exist $r, s \in M$ such that $p_U \in r$, $p_V \in s$ and $r^- \cap s^- = \emptyset$. By Lemma 2.157, $r \in U$, $s \in V$, so that U and V are not contacting. QED

LEMMA 2.160 *Let M_1 and M_2 be mereotopologies over weakly regular topological spaces, let $f : M_1 \simeq_C M_2$ be an isomorphism, and let U and V be contacting ultrafilters on M_1. Then $f(U)$ and $f(V)$ are contacting ultrafilters on M_2.*

Proof Almost immediate given the definability of $\leq$ in terms of C (Lemma 2.22). QED

LEMMA 2.161 *Let M_1 and M_2 be mereotopologies over regular topological spaces, such that $f : M_1 \simeq_C M_2$. Let $r \in M$, and let U be an ultrafilter on M_1 with $p_U \in r$. Then $p_{f(U)} \in f(r)$.*

Proof By Lemma 2.157, there exists $s \in U$ such that $p_U \in s$ and $s^- \subseteq r$, so that $s^- \cap (-r)^- = \emptyset$. Since f is also a Boolean algebra isomorphism, $f(s)^- \cap (-f(r))^- = \emptyset$, i.e. $f(s)^- \subseteq f(r)$. Since $f(s) \in f(U)$, $p_{f(U)} \in f(s)^- \subseteq f(r)$. QED

Proof [Theorem 2.154] Suppose that $f : M_1 \simeq_C M_2$. Define the map h by $h(p_U) = p_{f(U)}$, for U a compact ultrafilter on M_1. We show: (i) h is well-defined and 1–1, (ii) the domain of h is the whole of X_1 and the range of h is the whole of X_2, (iii) for all $r \in M_1$, $f(r) = h(r)$, and for all $s \in M_2$, $f^{-1}(s) = h^{-1}(s)$, and (iv) h and h^{-1} are continuous. To prove (i), let U, V be compact ultrafilters on M_1, both converging to p. By Lemma 2.160, the isomorphism f maps contacting ultrafilters to contacting ultrafilters. Hence, h is well-defined. Applying the same reasoning to f^{-1}, h is 1–1. To prove (ii), let $p \in X_1$. Then $\{r \in M_1 | p \in r\}$ is a proper filter on M_1, and by Proposition 2.155, this filter can be extended to an ultrafilter U on M_1. By Lemma 2.156, U converges to some point p_U. Since X_1 is Hausdorff, $p = p_U$. Thus, the domain of h is the whole of X_1. Similarly, if $q \in X_2$, we have an ultrafilter V on M_2 such that $q = p_V$. Thus $q = p_V = p_{f(f^{-1}(V))} = h(p_{f^{-1}(V)})$, so that the range of h is the whole of X_2. To prove (iii), let $p_V \in f(r)$ with V an ultrafilter on M_2. By Lemma 2.161, $p_{f^{-1}(V)} \in r$. Hence, $p_V = h(p_{f^{-1}(V)}) \in h(r)$. Conversely, let $p_V \in h(r)$. By the definition of h, $p_{f^{-1}(V)} \in r$, and by Lemma 2.161, $p_V \in f(r)$. Hence $f(r) = h(r)$. Now if $s \in M_2$, $f^{-1}(s) \in M_1$, so, applying the results just obtained to this set, we have $f^{-1}(s) = h^{-1}(h(f^{-1}(s))) = h^{-1}(f(f^{-1}(s))) = h^{-1}(s)$. (iv) Let $u \subseteq X_1$ be an open set. Since M_1 is a mereotopology, for each point $p \in u$,

there exists $r_p \in M_1$ with $p \in r_p \subseteq u$. Thus the set $\mathcal{U} = \{r_p \in M | p \in u\}$ satisfies $\bigcup \mathcal{U} = u$. Then $h(u) = h(\bigcup \mathcal{U}) = \bigcup_{r \in \mathcal{U}} h(r) = \bigcup_{r \in \mathcal{U}} f(r)$ is a union of open sets in X_2 and hence is itself an open set in X_2. Therefore, h^{-1} is continuous. By substituting h^{-1} and for h and repeating the argument, h is continuous. QED

7.2 Abstract models of geometrical mereotopological theories

We have shown that mereotopologies over certain classes of topological spaces can be characterized in terms of certain first-order sentences which they make true. But what of specific mereotopologies of interest—for instance, those defined over the open or closed plane? This is the topic we now address, based on the results of Pratt and Lemon, 1997.

We employ standard results on prime models: for details, see Chang and Keisler, 1990, Ch. 2. A structure $\mathfrak{A}$ is said to be a *prime model* if it is elementarily embeddable in any elementarily equivalent submodel. Prime models are considered the "simplest" or "smallest" models of their theories, a view which is justified by the following proposition (Chang and Keisler, 1990, Theorem 2.3.3). In the sequel, all signatures are silently assumed to be countable.

PROPOSITION 2.162 *Elementarily equivalent prime models are isomorphic.*

The following notion is closely related to that of primeness. A formula ϕ is said to be *complete with respect to* a theory T if, for all formulas θ having the same free variables of ϕ, exactly one of $T \models \phi \rightarrow \theta$ or $T \models \phi \rightarrow \neg\theta$ holds. A structure $\mathfrak{A}$ is said to be *atomic* if any n-tuple $\bar{a}$ in A satisfies a formula $\phi(\bar{x})$ in $\mathfrak{A}$ such that ϕ is complete with respect to $\mathrm{Th}(\mathfrak{A})$. We have the following standard result (see, for example, Chang and Keisler, 1990, Theorem 2.3.4).

PROPOSITION 2.163 *A structure is countable atomic if and only if it is a prime model.*

Recall the concepts of topologically complete formula and homogeneous mereotopology given in Definitions 2.51 and 2.90, respectively.

LEMMA 2.164 *Let M be a homogeneous mereotopology over a topological space X, and let Σ be a signature of topological primitives. If $\phi \in L_\Sigma$ is topologically complete in X over M, then ϕ is complete with respect to* $\mathrm{Th}_\Sigma(M)$.

Proof Immediate from Lemma 2.92. QED

Theorem 2.178 below is a partial converse of this result.

For the next theorem, recall that $\mathrm{ROQ}(\mathbb{S}^2)$ is the rational polygonal mereotopology over the closed plane, and that its $L_{c,\leq}$-theory is $\mathbf{T}_{c,\leq}$, the standard $L_{c,\leq}$-theory of closed plane mereotopology, which we axiomatized in Sec. 5. Recall further that $\psi^N_{\mathrm{c}^3}(\bar{z})$ is the $L_{c,\leq}$-formula stating that $\bar{z}$ forms a c^3-partition, employed in the proof of Theorem 2.85

THEOREM 2.165 *The mereotopology* $\mathrm{ROQ}(\mathbb{S}^2)$*, considered as a* $\{c, \leq\}$*-structure, is a prime model of* $\mathbf{T}_{c,\leq}$*. In fact, for any* N*, there exist formulas* $\gamma_1(\bar{z}), \ldots,$ $\gamma_K(\bar{z})$ *(with* K *depending on* N*), complete with respect to* $\mathbf{T}_{c,\leq}$*, such that*

$$\mathbf{T}_{c,\leq} \models \forall \bar{z}(\psi^N_{\mathrm{c}^3}(\bar{z}) \rightarrow (\gamma_1(\bar{z}) \vee \cdots \vee \gamma_K(\bar{z}))).$$

Proof The first part of the theorem is immediate from Theorem 2.85 and Lemma 2.164. For the second part, observe that, for a given N, there are only finitely many neighbourhood structures on an N-element c^3-partition, each one giving rise to a topologically complete formula of the form

$$\psi^N_{\mathrm{c}^3}(\bar{z}) \wedge \psi^{\bar{s}}_{+}(\bar{z}),$$

as described in the proof of Theorem 2.85. QED

Note that, by Lemma 2.38, $\mathrm{ROQ}(\mathbb{S}^2)$ and $\mathrm{ROQ}(\mathbb{R}^2)$ are the same $\{c, \leq\}$-structure, so we could replace $\mathbb{S}^2$ in Theorem 2.165 by $\mathbb{R}^2$.

Similarly, we have

THEOREM 2.166 *The mereotopologies* $\mathrm{ROQ}(\mathbb{R}^2)$ *and* $\mathrm{ROQ}(\mathbb{S}^2)$*, considered as* $\{C\}$*-structures, are prime models.*

Proof As for Theorem 2.165, but using Theorem 2.89 and Corollary 2.86, respectively. QED

Analogues of Theorem 2.166 hold in three dimensions, of course. For example, we have:

THEOREM 2.167 *The mereotopology* $\mathrm{ROQ}(\mathbb{R}^3)$ *is a prime model of the* L_C*-theory of* $\mathrm{ROP}(\mathbb{R}^3)$.

The proof strategy is essentially identical to the plane case, using Theorem 2.150. Note, however, that *much* more care is required to show that the topologically complete formulas identified in Theorem 2.150 are complete with respect to the L_C-theory of $\mathrm{ROP}(\mathbb{R}^3)$. We leave the details to the interested reader.

Returning to mereotopologies over $\mathbb{S}^2$, the question then arises as to whether $\mathrm{ROQ}(\mathbb{S}^2)$ is *strictly* simplest among countable models of $\mathbf{T}_{c,\leq}$, in that there are countable models of that theory not isomorphic to $\mathrm{ROQ}(\mathbb{S}^2)$. The answer

is: yes and no. Recall that a theory is said to be *ω-categorical* if it has exactly one countable model up to isomorphism. Recall also that a *type* in variables $\bar{x} = x_1, \ldots, x_n$ is a maximal consistent set of formulas whose free variables are among the $x_1, \ldots, x_n$, and that a theory T is said to *have* a type $\Phi(\bar{x})$ if $\Phi(\bar{x})$ is consistent with T. The following result is standard (see, for example, Chang and Keisler, 1990, Theorem 2.3.13).

PROPOSITION 2.168 *Let T be a complete theory. Then T is ω-categorical if and only if, for each n, T has only finitely many types in $x_1, \ldots, x_n$.*

THEOREM 2.169 *$\mathbf{T}_{c,\leq}$ is not ω-categorical.*

Proof By Proposition 2.168, it suffices to prove that $\mathbf{T}_{c,\leq}$ has infinitely many types in the single variable x. It is easy to see that, for every positive integer m, the formula $\psi_m(x)$

$$\exists z_1 \ldots \exists z_m \left(\bigwedge_{1 \leq i \leq m} (c(z_i) \wedge z_i \neq 0) \wedge \bigwedge_{1 \leq i < j \leq m} \neg c(z_i + z_j) \wedge x = \sum_{1 \leq i \leq m} z_i \right)$$

is satisfied in $\mathrm{ROQ}(\mathbb{S}^2)$ by all and only those regions having exactly m components. Hence, the $\psi_m(x)$ are all satisfied in $\mathrm{ROQ}(\mathbb{S}^2)$; so each can be extended to a type $\Gamma_m(x)$ of $\mathrm{Th}_{c,\leq}(\mathrm{ROQ}(\mathbb{S}^2))$. But the $\psi_m(x)$ are also pairwise mutually exclusive in $\mathbf{T}_{c,\leq}$; so no two of them can be extended to the same type. Hence, $\mathbf{T}_{c,\leq}$ has infinitely many types in x. QED

One the other hand, it turns out that $\mathbf{T}_{c,\leq}$ is *almost* countably categorical, in the following sense. Note that, since any model of $\mathbf{T}_{c,\leq}$ is a Boolean algebra interpreting the predicate c, we may employ the terminology introduced at the start of Sec. 4.1.

THEOREM 2.170 *All countable finitely decomposable models of $\mathbf{T}_{c,\leq}$ are isomorphic.*

Proof Let $\mathfrak{A} \models \mathbf{T}_{c,\leq}$ be finitely decomposable. By Claims 2.54 and 2.63, every tuple from A can be refined to a c^3-partition. Theorem 2.165 then implies that $\mathfrak{A}$ is prime. The result follows by Proposition 2.162. QED

The above results show that, while specific mereotopologies such as $\mathrm{ROS}(\mathbb{S}^2)$ cannot be characterized in terms of the first-order sentences which they make true, they *almost* can. Specifically, we have the following abstract characterization of the mereotopology $\mathrm{ROQ}(\mathbb{S}^2)$.

COROLLARY 2.171 *If $\mathfrak{A}$ is a countable, finitely decomposable model of Axioms* 1—8 *in Sec. 5.1, then $\mathfrak{A}$ is isomorphic* (*as a $\{c, \leq\}$-structure*) *to the mereotopology* $\mathrm{ROQ}(\mathbb{S}^2)$.

Proof Theorem 2.170 and the fact that, by Theorem 2.101, any finitely decomposable model $\mathfrak{A}$ of Axioms 1—8 is elementarily equivalent to $\mathrm{ROQ}(\mathbb{S}^2)$. QED

7.3 Loose ends

We end this section with some matters touched on earlier in this chapter. We continue to assume all signatures to be countable. The following proposition is a special case of the Löwenheim-Skolem Theorem (see, for example, Hodges, 1993, p. 90).

PROPOSITION 2.172 *Let $\mathfrak{A}$ be a Σ-structure and Z a countable subset of A. Then $\mathfrak{A}$ has a countable elementary submodel whose domain includes Z.*

Recall that a topological space X is said to be *second countable* if its topology has a countable basis.

LEMMA 2.173 *Let M be a mereotopology over a compact, second-countable, Hausdorff space X, and let $P \subseteq M$ be countable. Then there is a countable mereotopology Q over X such that $P \subseteq Q$ and $Q \preceq M$.*

Proof We construct a countable subset $P' \subseteq M$ such that, for all $p \in o \subseteq X$ with o open, there exists $r \in P'$ such that $p \in r \subseteq o$. The lemma then follows from Proposition 2.172 by putting $\mathfrak{A} = M$ and $Z = P \cup P'$. Let B be a countable basis for the topology on X. For any $b, c \in B$ with $b^- \subseteq c$, take a cover of b^- by elements $s \in M$ such that $s \subseteq c$ (possible because M is a mereotopology), choose a finite subcover (possible because X is compact), and let $r_{b,c}$ be the sum, in M, of the elements of this finite subcover. Certainly, $b \subseteq r_{b,c} \subseteq c^-$. Let $P' = \{r_{b,c} \mid b, c \in B, b^- \subseteq c\}$. Since X is normal, for all $p \in o \subseteq X$ with o open, we can find $b, c \in B$ with $p \in b$, $b^- \subseteq c$ and $c^- \subseteq o$. But then $p \in r_{b,c} \subseteq o$ as required. QED

Note that Lemma 2.173 holds for all (countable) signatures.

We may now derive the promised strengthening of Corollary 2.118.

COROLLARY 2.174 *All splittable, finitely decomposable mereotopologies over $\mathbb{S}^2$ with curve-selection have the same L_Σ-theory for any topological signature Σ.*

Proof Let M_1, M_2 be two such mereotopologies. Extend the signature Σ if necessary so that it contains the predicates C, c and $\leq$, and expand M_1 and M_2 by interpreting these predicates in the normal way. By Lemma 2.173, let Q_i be a countable mereotopology over $\mathbb{S}^2$ such that $Q_i \preceq_\Sigma M_i$, for $i = 1, 2$. Thus, Q_1 and Q_2 are splittable, finitely decomposable mereotopologies over $\mathbb{S}^2$ having curve-selection. By Corollary 2.118, $Q_1 \equiv_{c,\leq} Q_2$. By Theorem 2.170, $Q_1 \simeq_{c,\leq} Q_2$. By Lemma 2.49, $Q_1 \simeq_C Q_2$. By Theorem 2.154, there is a

homeomorphism mapping Q_1 onto Q_2. Finally, by Lemma 2.153, $Q_1 \simeq_\Sigma Q_2$, whence $M_1 \equiv_\Sigma M_2$. QED

Recall from Definition 2.97 that, if Σ is a signature of topological primitives, $\mathbf{T}_\Sigma$ denotes $\mathrm{Th}_\Sigma(\mathrm{ROS}(\mathbb{S}^2))$. By Corollary 2.174, $\mathbf{T}_\Sigma$ is the L_Σ-theory of any splittable, finitely decomposable mereotopology over $\mathbb{S}^2$ having curve-selection. This justifies our decision to call it the *standard* L_Σ-theory of closed plane mereotopology.

Theorem 2.170 now has the following corollaries.

COROLLARY 2.175 *Let M be a countable, finitely decomposable mereotopology over a locally connected, compact, Hausdorff space X, such that $\mathrm{Th}_C(M) = \mathbf{T}_C$. Then there is a homeomorphism $h : X \leftrightarrow \mathbb{S}^2$ taking M to $\mathrm{ROQ}(\mathbb{S}^2)$.*

Proof By Lemmas 2.22 and 2.27, $\mathrm{Th}_{C,c,\leq}(M) = \mathbf{T}_{C,c,\leq}$. By Theorem 2.170, $M \simeq_{c,\leq} \mathrm{ROQ}(\mathbb{S}^2)$. But, by Lemma 2.49, $\mathbf{T}_{C,c,\leq}$ contains a formula defining C explicitly in terms of c and $\leq$. Hence $M \simeq_C \mathrm{ROQ}(\mathbb{S}^2)$. Now apply Theorem 2.154. QED

COROLLARY 2.176 *Let M be a finitely decomposable mereotopology over a locally connected, second countable, compact, Hausdorff space X, such that $\mathrm{Th}_C(M) = \mathbf{T}_C$. Then X is homeophorphic to $\mathbb{S}^2$.*

Proof Apply Lemma 2.173 to obtain a countable mereotopology Q over X with $Q \preceq M$ and proceed as for Corollary 2.175. QED

We remark that there is no prospect of removing the compactness condition from the above corollaries. For example, let p_π be, say, the point of $\mathbb{S}^2$ with coordinates $(0, \pi)$, and consider the topological space $X = \mathbb{S}^2 \setminus \{p_\pi\}$ and the mereotopology M over X given by $M = \{r \setminus \{p_\pi\} \mid r \in \mathrm{ROQ}(X)\}$. Then $\mathrm{ROQ}(\mathbb{S}^2) \simeq_{C,c,\leq} M$; but $\mathbb{S}^2$ and X are not homeomorphic.

A further consequence of Theorem 2.154 is the promised partial converse of Lemma 2.164. We require the following fact about prime models.

LEMMA 2.177 *Let $\mathfrak{A}$ be a countable, atomic model and let $\bar{a}$,$\bar{b}$ be tuples from A which satisfy the same formulas in $\mathfrak{A}$. Then there is an automorphism of $\mathfrak{A}$ taking $\bar{a}$ to $\bar{b}$.*

Proof Almost immediate from Proposition 2.162, by adding a tuple of individual constants to stand alternatively for $\bar{a}$ and $\bar{b}$. QED

THEOREM 2.178 *Let M be a mereotopology over a compact, second-countable Hausdorff space X, and let Σ be a signature of topological primitives such that*

C (contact) is first-order definable over M. If every tuple from M satisfies an L_Σ-formula which is complete with respect to Th_Σ, then that L_Σ-formula is topologically complete in M over X.

Proof Let ϕ be complete with respect to $\mathrm{Th}_\Sigma(M)$, and suppose that $M \models \phi[\bar{r}]$, $M \models \phi[\bar{s}]$. We must show that $\bar{r}$ and $\bar{s}$ are similarly situated in X. By Lemma 2.173, let M' be a countable mereotopology over X containing the tuples $\bar{r}$ and $\bar{s}$, such that $M' \preceq M$. Thus, M' is countable and atomic, and ϕ is a complete formula with respect to $\mathrm{Th}_\Sigma(M')$ satisfied by both $\bar{r}$ and $\bar{s}$. By Lemma 2.177, there exists an automorphism $f : M' \simeq_\Sigma M'$ such that $f(\bar{r}) = \bar{s}$. Then, by Theorem 2.154, there is a homeomorphism $h : X \to X$ taking $\bar{r}$ to $\bar{s}$. QED

Lemma 2.164 and Theorem 2.178 establish the close connection between the notions of topological completeness with respect to a topological space and completeness with respect to a mereotopological theory.

8. Philosophical considerations

The earliest modern work on region-based theories of space is that of Whitehead and de Laguna (Whitehead, 1919; Whitehead, 1920; Whitehead, 1929; de Laguna, 1922a; de Laguna, 1922b; de Laguna, 1922c). Both authors propose a system of postulates governing a small collection of primitive spatial relations, together with reconstructions of familiar spatial concepts in terms of those relations. The postulates serve implicitly to define the primitive relations they constrain (and perhaps the domain of entities over which they quantify), while the reconstructions of familiar spatial concepts connect the whole system to the data of spatial experience. To be sure, both Whitehead and de Laguna motivate their postulates by providing informal interpretations for their respective spatial primitives. Thus, for example, Whitehead illustrates his relation of *extensive connection* (as he calls it) using diagrams suggesting that two regions are extensively connected just in case their topological closures share a point in common (this is the interpretation given to the binary predicate C in this chapter). However, such explanations are intended only as a heuristic guide. Officially, spatial primitives acquire their content solely from the entire system postulates in which they participate. Primitives, by definition, are not explicitly definable.

The inspiration for such systems was presumably the axiomatic treatment of geometry found in Euclid (and latterly Hilbert); and the motivation for carrying out the procedure on a purely region-based footing seems, for both authors, to have been a certain disquiet about the empirical distance between the concept of a point as a primitive geometrical entity and the character of everyday spatial experience. The great difficulty of this approach, of course, is the problem of evaluating the system of postulates and conceptual reconstructions proposed.

Whitehead's system has thirty-one postulates (or *assumptions*, in Whitehead's terminology) and a similar number of definitions. De Laguna's system, though far tidier, is also hardly self-evident. The only obvious sources of justification for such systems are their ability to chime with our pre-theoretic intuition and their eventual integration into a larger, empirically successful, physical theory. Neither source is very satisfactory. On the one hand, as we have seen in this chapter, almost any collection of spatial primitives enables us to write down propositions on which pre-theoretic intuition cannot be expected to return a reliable verdict. On the other hand, although empirical confirmation of a general physical theory must provide some support for the account of space it contains, the size of the undertaking and the difficulty of assigning credit when theories perform well (or blame when they perform badly) means that there is little practical prospect of any such justification for such systems of postulates and conceptual reconstructions.

An alternative approach to developing a region-based theory of space is illustrated by Tarski's *Geometry of Solids* (Tarski, 1956). Tarski too develops a geometry in which regions, not points, are the primitive objects; however, in contrast to Whitehead and de Laguna, he does not build his theory by writing a collection of plausible-looking, but unprovable, axioms. Rather, beginning with the familiar model of space as $\mathbb{R}^3$, he considers a formal language whose variables range over the set of of spheres in $\mathbb{R}^3$ (defined in the standard way), and whose sole non-logical constant is the part-whole relation (again, defined in the standard way). Because the "primitives" in Tarski's Geometry of Solids are well-defined mathematical objects and relations, the question of what postulates they satisfy is a well-defined mathematical problem, not a matter for intuition or experiment. And because many familiar spatial concepts have rational reconstructions in terms of the standard model, the question of how, if at all, these concepts can be expressed using formulas of Tarski's language is again a purely mathematical affair. Having thus specified the structure under consideration and the language used to describe it, Tarski then goes on to examine the kinds of logical issues that should by now be familiar to us. In fact, Tarski obtains a system of axioms (in higher-order logic) for which the standard Euclidean interpretation is, up to isomorphism, the only model.

This alternative approach is, in contrast to the "postulationist" strategy of Whitehead and de Laguna, conservative and rationalist: conservative, because no attempt is made to build systems of axioms and definitions from the ground up; rationalist, because the appropriateness of the resulting region-based theories is secured by means of their logical relations to point-based models whose usefulness as representations of the space we inhabit—at least approximately and for mesoscopic objects—is anyway beyond doubt. It is this approach that we have taken in this chapter. Latterly, region-based theories of space have increased in popularity, following the seminal work of Clarke, 1981, Clarke,

1985, Biacino and Gerla, 1991, Randell et al., 1992, Gotts et al., 1996 and Renz and Nebel, 1997. One reason for this resurgence of interest, particularly within the AI community, is the requirement to quantify over spatial regions without leaving the realm of first-order logic. The technology of theorem-proving for first-order logic is more highly developed than for higher-order logics; and, more generally, formalisms with limited expressive power enjoy a premium in AI if they give rise to entailment and satisfiability problems which have (theoretically or practically) efficient algorithmic solutions. Insofar as the study of region-based theories of space is motivated by computational considerations, the best approach to developing and analysing such theories is surely that of Tarski, not that of Whitehead.

These matters notwithstanding, the principal outcome of the investigation undertaken here is just how much information it gives us about the possibilities for developing a truly region-based theory of space, along the lines apparently envisaged by Whitehead and de Laguna. Consider, for example, the issue of the "correct" set of postulates. True, Examples 2.17 and 2.18 show that different mereotopologies defined over the spaces $\mathrm{RO}(\mathbb{R}^2)$ indeed have different first-order theories. Nevertheless, the discussion of Sec. 5 shows that the choices on offer are much more limited than these examples might initially lead one to suppose. In particular, all finitely decomposable, splittable mereotopologies over $\mathbb{S}^2$ having curve-selection have identical L_Σ-theories, for any signature of topological primitives. We proposed that this common L_Σ-theory should therefore be regarded as standard.

Or take again the issue of reconstructing familiar spatial concepts in terms of a chosen collection of primitives. We have seen that first-order topological languages interpreted over well-behaved mereotopologies have surprising—but not unlimited—expressive power. In particular, we provided formulas expressing a variety of familiar spatial relationships (as defined by their usual point-based definitions, of course) over a wide range of mereotopologies. In addition, we showed that the first-order language $L_{c,\leq}$ is sufficiently expressive that every tuple of polygons in $\mathbb{S}^2$ can be characterized up to similar situation by one of its formulas, and that the first-order language L_C is sufficiently expressive that every tuple of polygons in $\mathbb{R}^2$ and every tuple of polyhedra in $\mathbb{R}^3$ can be characterized up to similar situation by one of its formulas.

Most striking of all, however, is what the foregoing analysis tells us about the view of space to which any first-order mereotopological theory commits us. While almost all interesting mereotopologies have first-order theories which are not categorical in any infinite cardinal, we nevertheless showed that the plane mereotopology $\mathrm{ROQ}(\mathbb{S}^2)$ and the spatial mereotopology $\mathrm{ROQ}(\mathbb{S}^3)$ are prime models of their first-order theories over standard signatures of topological primitives. We further showed that $\mathrm{ROQ}(\mathbb{S}^2)$ is, up to isomorphism, the only countable, finitely decomposable model of its $L_{c,\leq}$-theory; and we

remarked that a corresponding observation—albeit with a more complex version of the finite decomposability condition—must apply in three dimensions as well. Finally, we showed that mereotopologies over compact, Hausdorff spaces, regarded as structures interpreting suitably rich topological signatures, determine their underlying spaces up to homeomorphism. In conclusion, the logical possibilities for region-based topological theories of space are more constrained than their earliest proponents might perhaps have thought.

Acknowledgment

The author wishes to thank Mr. Aled Griffiths for help in preparing the manuscript.

References

Alexander, J.W. (1924a). An example of a simply connected surface bounding a region which is not simply connected. *Proceedings of the National Academy of Sciences of the United States of America*, 10(1):8–10.

Alexander, J.W. (1924b). On the subdivision of 3-space by a polyhedron. *Proceedings of the National Academy of Sciences of the United States of America*, 10(1):6–8.

Biacino, L. and Gerla, G. (1991). Connection structures. *Notre Dame Journal of Formal Logic*, 32(2):242–247.

Bochnak, J., Coste, M., and Roy, M.-F. (1998). *Real Algebraic Geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*, 3, *Folge*. Springer Berlin.

Cantone, D. and Cutello, V. (1994). Decision algorithms for elementary topology I. Topological syllogistics with set and map constructs, connectedness and cardinailty composition. *Communications on Pure and Applied Mathematics*, XLVII:1197–1217.

Chang, C.C. and Keisler, H.J. (1990). *Model Theory*. North Holland, Amsterdam, 3rd edition.

Clarke, B.L. (1981). A calculus of individuals based on "connection". *Notre Dame Journal of Formal Logic*, 22(3):204–218.

Clarke, B.L. (1985). Individuals and points. *Notre Dame Journal of Formal Logic*, 26(1):61–75.

Davis, E., Gotts, N.M., and Cohn, A.G. (1999). Constraint networks of topological relations and convexity. *Constraints*, 4(3):241–280.

de Laguna, T. (1922a). The nature of space–I. *The Journal of Philosophy*, 19:393–407.

de Laguna, T. (1922b). The nature of space–II. *The Journal of Philosophy*, 19:421–440.

de Laguna, T. (1922c). Point, line, and surface as sets of solids. *The Journal of Philosophy*, 19:449–461.

Diestel, R. (1991). *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer, New York.

Dimov, G. and Vakarelov, D. (2006). Contact algebras and relation-based theory of space: a proximity approach I. *Fundamenta Informaticae*, 74:209–249.

Dornheim, C. (1998). Undecidability of plane polygonal mereotopology. In Cohn, A.G., Schubert, L.K., and Schubert, S.C., editors, *Principles of Knowledge Representation and Reasoning: Proc. 6th International Conference (KR '98)*, pages 342–353. Morgan Kaufmann, San Francisco.

Düntsch, I. and Winter, M. (2005). A representation theorem for Boolean contact algebras. *Theoretical Computer Science*, 347:498–512.

Egenhofer, M.J. (1991). Reasoning about binary topological relations. In Günther, O. and Schek, H.-J., editors, *Advances in Spatial databases, SSD '91 Proceedings*, volume 525 of *Lecture Notes in Computer Science*, pages 143–160. Springer, Berlin.

Erlich, G., Even, S., and Tarjan, R.E. (1976). Intersection graphs of curves in the plane. *Journal of Combinatorial Theory B*, 21:8–20.

Gotts, N.M., Gooday, J.M., and Cohn, A.G. (1996). A connection based approach to commonsense topological description and reasoning. *Monist*, 79(1): 51–75.

Hodges, W. (1993). *Model Theory*, volume 42 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge.

Kelley, J.L. (1955). *General Topology*, volume 27 of *Graduate Texts in Mathematics*. Springer, New York.

Koppelberg, S. (1989). *Handbook of Boolean Algebras*, volume 1. North-Holland, Amsterdam.

Kratochvíl, J. (1988). String graphs II: Recognizing string graphs is NP-hard. *Journal of Combinatorial Theory B*, 52:67–78.

Kuijpers, B., Paradaens, J., and van den Bussche, J. (1995). Lossless representation of topological spatial data. In Egenhofer, M.J. and Herring, J.R., editors, *Proc. 4th International Symposium on Large Spatial Databases*, volume 951 of *Lecture Notes in Computer Science*, pages 1–13. Springer, Berlin.

Massey, W.S. (1967). *Algebraic Topology : an Introduction*. Harcourt, Brace & World, New York.

Moise, E. (1977). *Geometric Topology in Dimensions 2 and 3*, volume 47 of *Graduate Texts in Mathematics*. Springer, New York.

Newman, M.H.A. (1964). *Elements of the Topology of Plane Sets of Points*. Cambridge University Press, Cambridge.

Nutt, W. (1999). On the translation of qualitative spatial reasoning problems into modal logics. In Burgard, W., Christaller, T., and Cremers, A., editors,

Advances in Artificial Intelligence, Proc. 23rd Annual German Conference on Artificial Intelligence, KI'99, volume 1701 of *Lecture Notes in Computer Science*, pages 113–124, Springer, Berlin.

Papadimitriou, C.H., Suciu, D., and Vianu, V. (1999). Topological queries in spatial databases. *Journal of Computer and System Sciences*, 58(1):29–53.

Pratt, I. (1999). First-order spatial representation languages with convexity. *Spatial Cognition and Computation*, 1:181–204.

Pratt, I. and Lemon, O. (1997). Ontologies for plane, polygonal mereotopology. *Notre Dame Journal of Formal Logic*, 38(2):225–245.

Pratt, I. and Schoop, D. (1998). A complete axiom system for polygonal mereotopology of the real plane. *Journal of Philosophical Logic*, 27(6):621–658.

Pratt, I. and Schoop, D. (2000). Expressivity in polygonal, plane mereotopology. *Journal of Symbolic Logic*, 65(2):822–838.

Pratt, I. and Schoop, D. (2002). Elementary polyhedral mereotopology. *Journal of Philosophical Logic*, 31:461–498.

Pratt-Hartmann, I. (2002). A topological constraint language with component counting. *Journal of Applied Non-Classical Logics*, 12(3–4):441–467.

Randell, D.A., Cui, Z., and Cohn, A.G. (1992). A spatial logic based on regions and connection. In Nebel, B., Rich, C., and Swartout, W., editors, *Principles of Knowledge Representation and Reasoning: Proc. 3rd International Conference (KR '92)*, pages 165–176. Morgan Kaufmann, San Francisco.

Renz, J. and Nebel, B. (1997). On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the region connection calculus. In *Proc. 15th International Joint Conference on Artificial Intelligence (IJCAI 97)*, pages 522–527. Morgan Kaufmann, San Francisco.

Roeper, P. (1997). Region-based topology. *Journal of Philosophical Logic*, 26:251–309.

Rudin, M.E. (1958). An unshellable triangulation of a tetrahedron. *Bulletin of the American Mathematical Society*, 64:90–91.

Schaefer, M., Sedgwick, E., and Štefankovič, D. (2003). Recognizing string graphs in NP. *Journal of Computer and System Sciences*, 67:365–380.

Schaefer, M. and Štefankovič, D. (2004). Decidability of string graphs. *Journal of Computer and System Sciences*, 68:319–334.

Schoop, D. (1999). *A Logical Approach to Mereotopology*. PhD thesis, Department of Computer Science, University of Manchester, Manchester, England.

Simons, P. (1987). *Parts: a Study in Ontology*. Clarendon Press, Oxford.

Steen, L.A. and Seebach, J.A. (1995). *Counterexamples in Topology*. Dover Publications, New York.

Tarski, A. (1956). Foundations of the geometry of solids. In *Logic, Semantics and Metamathematics*, pages 24–29. Clarendon Press, Oxford.

van den Dries, L. (1998). *Tame Topology and O-Minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge.

Whitehead, A.N. (1919). *An Enquiry Concerning the Principles of Natural Knowledge*. Cambridge University Press, Cambridge.

Whitehead, A.N. (1920). *The Concept of Nature*. Cambridge University Press, Cambridge.

Whitehead, A.N. (1929). *Process and Reality*. The MacMillan Company, New York.

Wilson, R.J. (1979). *Introduction to Graph Theory*. Longman, London.

Chapter 3

AXIOMS, ALGEBRAS AND TOPOLOGY

Brandon Bennett
University of Leeds

Ivo Düntsch
Brock University

Second Reader

István Németi
Rényi Mathematical Research Institute, Budapest

1. Introduction

This work explores the interconnections between a number of different perspectives on the formalisation of space. We begin with an informal discussion of the intuitions that motivate these formal representations.

1.1 Axioms *vs* algebras

Axiomatic theories provide a very general means for specifying the logical properties of formal concepts. From the axiomatic point of view, it is symbolic formulae and the logical relations between them—especially the entailment relation—that form the primary subject of interest. The vocabulary of concepts of any theory can be interpreted in terms of a domain of entities, which exemplify properties, relations and functional mappings corresponding to the formal symbols of the theory. Moreover, by interpreting logical operations as functions of these semantic denotations, such an interpretation enables us to evaluate the truth of any logical formula built from these symbols. An interpretation is said to satisfy, or be a *model* of a theory, if all the axioms of the theory are true according to this evaluation.

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 99–159.
© 2007 *Springer.*

In general, an axiomatic theory can have many different models exhibiting diverse structural properties. However, in formulating a logical theory, we will normally be interested in characterising a particular domain and a number of particular properties, relations and/or functions that describe the structure of that domain; or, more generally, we may wish to characterise a family of domains that exhibit common structural features, and which can be described by the same conceptual vocabulary.

From the *algebraic* perspective, it is the domain of objects and its structure that form the primary subject of investigation. Here again, we may be interested in a specific set of objects and its structure, or a family of object sets exemplifying shared structural features. And the nature of the structure will be described in terms of properties, relations and functions of the objects. To specify a particular structure or family of structures, one will normally give an axiomatic theory formulated in terms of this vocabulary, such that the algebraic structures under investigation may be identified with the models of the theory.

Hence, axiom systems and algebras are intimately related and complementary views of a conceptual system. The axiomatic viewpoint characterises the meanings of concepts in terms of propositions involving those concepts, whereas the algebraic viewpoint exemplifies these meanings in terms of a set of objects and mappings among them. Moreover, the models of axiomatic theories can be regarded as algebras, and conversely algebras may be characterised by axiomatic theories.

Having said this, the two perspectives lead to different emphasis in the way a conceptual system is articulated. If one starts from axiomatic propositions, one tends to focus on relational concepts (formalised as predicates), whereas, if one starts from objects and structures, the focus tends to be on functional concepts corresponding to mappings between the objects. Indeed, the term 'algebra' is sometimes reserved for structures that may be characterised without employing any relational concept apart from the logical equality relation. And the most typical algebras are those specified purely by means of universally quantified equations holding between functional terms.

1.2 Representing space

1.2.1 Classical approaches. Our modern appreciation of space is very much conditioned by mathematical representations. In particular, the insights into spatial structure given to us by Euclid and Descartes are deeply ingrained in our understanding.

Euclid described space in terms several distinct categories of geometrical object. These include *points*, *lines* and *surfaces* as well as *angles* and *plane figures*. These entities may be said to satisfy a number of basic properties (e.g. lines may be *straight* and surfaces may be *planar*) and relationships (e.g. a

point may be *incident* in a line or surface, two lines may meet at a point or be inclined at an angle). The nature of space was then characterised by postulates involving these basic concepts, which were originally stated in ordinary language. Euclid proceeded to define many further concepts (such as different types of geometrical figure) in terms of the basic vocabulary.

From Descartes came a numerical interpretation of space, with points in an n-dimensional space being associated with n-tuples of numerical values. According to this Cartesian model, the basic elements of space are points. Their structure and properties can be axiomatised in terms of the metrical relation of equidistance (see e.g. Tarski, 1959; Tarski and Givant, 1999), and interpreted in terms of numerical coordinates.

If points are taken as the primary constituents of the universe, lines and regions have a derivative status. Two distinct points determine a line, and polygonal figures can be represented by a sequence of their vertex points. To get a more general notion of 'region' we need to refer to more or less arbitrary collections of points. This is the representational perspective of classical point-set topology.

1.2.2 Region-based approaches. Although the point-based analysis has become the dominant approach to spatial representation, there are a number of motivations for taking an alternative view, in which *extended regions* are considered as the primary spatial entities.

An early exponent of this approach was Alfred North Whitehead, who shared with Bertrand Russell the view that an adequate theory of nature should be founded on an analysis of *sense data*, and that elements of perception can be the only referents of truly primitive terms. On this basis, Whitehead in his book *Concept of Nature* (Whitehead, 1920) argued that extended regions are more fundamental than points: whereas regions may be perceived as the spatial correlates of colour patches in the visual field, points cannot be perceived directly but are only constructed by cognitive abstraction.

This motivation shares some common ground with that of Stanislaw Leśniewski, who also wanted to bring the theoretical analysis of the world more closely in line with phenomenological conceptions of reality, and believed that perceiving the integrity of extended objects is basic to our interpretation of the world. *Mereology*, a formal theory of the part-whole relation was originally presented by Leśniewski, 1931 in his own logical calculus, which he called *Ontology*.

Whitehead identified the relation of *connection* between two spatial or spatio-temporal regions as of particular importance to the phenomenological description of reality. In further work he attempted to use this concept as the fundamental primitive in a logical theory of space and time. A formal theory of this relation was presented in Whitehead, 1929. (This was subsequently

found to be inconsistent and was posthumously corrected in a second edition, Whitehead, 1978.)

The earliest completely rigorous and fully formalised theory of space where regions are the basic entity is the axiomatisation of a *Geometry of Solids* that was given by Tarski, 1929. Subsequently, a number of other formalisations have been developed. These include the *Calculus of Individuals* proposed by Leonard and Goodman, 1940 and Goodman, 1951, which is close to Leśniewski's mereology, and the spatial theories of Clarke, 1981 and Clarke, 1985, which are based on Whitehead's connection relation.

More recently, region-based theories have attracted attention from researchers working on Knowledge Representation for Artificial Intelligence (AI) systems. The so-called *Region Connection Calculus* (Randell et al., 1992b) is a 1st-order formalism based on the connection relation and is a modification of Clarke's theory. AI researchers are motivated to study such representations by a belief that they may be useful as a vehicle for automating certain human-like spatial reasoning capabilities. From this point of view it has been argued that the region-based approach is closer to the natural human conceptualisation of space. In the context of describing and reasoning about spatial situations in natural language, it is common to refer directly to regions and the relations between them, rather than referring to points and sets of points. Therefore, treating regions as basic entities in a formal language can in many cases allow simpler representation of high-level human-like spatial descriptions.

1.2.3 Interdefinability of regions and points. Despite the difference in perspective, several formal results show that region and point-based conceptualisations are in fact interdefinable, given a sufficiently rich formal apparatus. Whitehead himself had noted that by considering classes of regions, points can be defined as infinite sets of nested regions which converge to a point. Pratt and Lemon, 1997 and Pratt-Hartmann, 2001 showed that any sufficiently strong axiomatisation of the polygonal regions of a plane can be interpreted in terms of the classical point-based model of the Euclidean plane. So in some sense the region-based theory is not 'ontologically simpler' in its existential commitments.

Later in this paper we shall adopt a similar approach, and by identifying a point with the sets of all regions region to which it belongs, we shall show that even much weaker and more general region-based theories can be interpreted in terms of point sets. Hence, the point and region based approaches should not be regarded as mutually exclusive, but rather as complementary perspectives.

Nevertheless, we believe that region-based theories deserve more attention than has traditionally been paid them, and that for certain purposes they have clear advantages.

There is an argument that regions are actually more powerful and flexible than points as a starting point for spatial theories. Pratt-Hartmann, 2001 shows that if we have a domain of regions plus a spatial language with sufficient (in fact rather low) expressive power, we implicitly determine a corresponding domain of points. Roughly speaking, this is done as follows: within the region-based theory we can specify pairs of regions that have a unique 'point' of contact. Thus, relations among points can be recast in the guise of formulae which refer to these region pairs. In this way points are implicitly definable from regions by 1st-order means; whereas, if we start with points as the basic entities, the definition of regions requires higher order axioms (unless we arbitrarily restrict the geometrical complexity of regions).

1.3 Alternative logical formalisms

We have seen that the representation of space allows alternative views that invert the perspective of the orthodox picture. The same is also true for the mode of application of formal representations themselves. 1st-order logic provides a standard alignment of syntactic and conceptual categories. Specifically, the basic nominal symbols of the formal language refer to what are considered to be the primitive entities of the 'domain' of a theory, while formal predicates correspond to properties and relations that hold among those entities. This alignment is widely held to be 'natural', in that it seems to accord in some respects with the syntactic expression of semantics found in natural languages. However, this intuition is difficult to establish conclusively. Moreover, by altering the correspondence between syntactic and conceptual categories, one may obtain alternative calculi that also have a meaningful interpretation and proof theory. For instance, one could interpret syntactically basic symbols as denoting 'properties', and represent 'individuals' formally as predicates (the extension of the predicate being the set of properties satisfied by the corresponding individual).

However, since relations between objects are not in general reducible to properties of individual objects, a much more powerful abstraction is obtained by taking *relations* as the basic entities of a formal system. This approach was first formalised in an algebraic framework by Tarski, 1941 and relation algebras are now a well-established alternative to standard 1st-order formalisms (Tarski and Givant, 1987; Andréka et al., 2001).

From the point of view of logic and computation, there are significant advantages in treating relations as basic entities. In particular this mode of representation allows quantifier-free formalisation of many properties and inference patters, which would otherwise require quantification. This is one of

the main themes of algebraic logic, as it is elaborated in Andréka et al., 2001; Németi, 1991 and Ahmed, 2004.

As we will be concerned with spatial representation based on the 'contact' relation, relation algebras are a natural system within which to formulate theories of this kind.

1.4 Structure of the chapter

The organisation of this chapter is as follows. In Sec. 2 we shall present some basic formal structures and notations that will be used to develop the theory. These include, Boolean algebras, relation algebras, topological spaces and proximity spaces. Sec. 3 introduces the spatial *contact* relation, which is the primary focus of our investigation. We consider the fundamental axioms satisfied by a contact relation and give standard interpretations of contact in terms of point-set topology. We then see how the basic properties of this relation can be described in both 1st-order and relation algebraic calculi.

In Sec. 4 we introduce Boolean Contact Algebras, which are Boolean algebras supplemented with a contact relation satisfying appropriate general axioms. Additional axioms are also considered, which characterise further properties of contact that are exhibited under typical spatial interpretations. We give representation theorems for the general class of Boolean Contact Algebras in terms of both topological spaces and proximity spaces, and give more specific representation systems for algebras satisfying additional axioms. These theorems make concrete the correspondence between the relational approaches which focus on axiomatic properties of the contact relation, and the more well-known models of space in terms of point-set topology.

Sec. 5 looks at some other well known approaches to formalising topological relationships, in particular the Region Connection Calculus (Randell et al., 1992b) and the 9-intersection model (Egenhofer and Franzosa, 1991). Sec. 6 considers the problem of reasoning with topological relations. The methods presented are: compositional reasoning, equational reasoning, encoding into modal logic and a relation algebraic proof theory. Sec. 7 concludes the chapter with a consideration of the correspondences that have been established between different modes of formalising topological information, and of ongoing and future developments in this area.

2. Preliminary definitions and notation

In this section we give definitions and key properties of the basic formal structures that will underpin our analysis. We start with *Boolean algebras with operators*, which provide an extremely general framework for studying structured domains of objects. Two important special cases of these algebras are considered: modal algebras, and relation algebras.

2.1 Boolean algebras

We assume that the reader has some familiarity with Boolean algebras (BAs), and here only revise the basic details and notation. Our standard reference for BAs is Koppelberg, 1989. Our signature for a BA will be $\langle B, \cdot, +, -, \mathbf{0}, \mathbf{1}\rangle$. We will usually refer to an algebra by its base set (in this case B).

DEFINITION 3.1 *Boolean algebra concepts and notations:*

i) *For all $a, b \in B$, $a \leq b$ holds iff $a + b = b$.*

ii) *If $A \subset B$, then $\sum_B A$ denotes the least upper bound of A relative to the $\leq$ ordering of B. If A is infinite this does not necessarily exist. Where the relevant algebra is clear, we may write simply $\sum A$.*

iii) *If A is a subalgebra of B, we denote this by $A \leq B$.*

iv) *The set of non-zero elements of B is denoted by B^+.*

v) *If M is a subset of B^+, then M is* dense in *B, iff*

$$(\forall b \in B^+)(\exists a \in M)\ a \leq b\ .$$

vi) *An* atom of *B is an element $a \in B^+$ such that*

$$(\forall c)[c \leq a \rightarrow (c = \mathbf{0} \vee c = a)]\ .$$

vii) *The set of atoms of B will be written as* $\mathrm{At}(B)$.

viii) *B is* atomic, *iff* $\mathrm{At}(B)$ *is dense in B^+.*

ix) *If $f : B \rightarrow B$ is a mapping, then its* dual *is the mapping $f^\partial : B \rightarrow B$ defined by $f^\partial(x) = -f(-x)$.*

In general, a BA will contain elements corresponding to the meet and join of any *finite* subset of its domain. A BA is called *complete*, if arbitrary joins and meets exist. The *completion of B* is the smallest complete BA A which contains B as a dense subalgebra. It is well known that each B has a completion which is unique up to isomorphisms.

DEFINITION 3.2 *An* ultrafilter *is a subset F of B such that:*

i) *If $x \in F$, $y \in B$ and $x \leq y$, then $y \in F$.*

ii) *If $x, y \in F$, then $x \cdot y \in F$.*

iii) *$x \in F$ if and only if $-x \notin F$.*

The set of all ultrafilters of B will be denoted by Ult(B).

Ultrafilters are often employed as a means to represent 'point-like' entities that are implicit in the structure of a BA. From a purely algebraic point of view, the elements of a BA are abstract entities with no sub-structure. However, the elements may be, and often are, intended to correspond to composite objects (e.g. sets or spatial regions). Thus, as we shall see later, the elements of a BA are often *interpreted* as point sets in some space (e.g. a topological space). In such a context, an ultrafilter can usually be thought of a set of all those elements of a BA that contain some particular point in the space over which the algebra is interpreted.

Perhaps the simplest example is the BA X^* whose elements are (interpreted as) all subsets of the set X (with the Boolean operations having their standard set-theoretic interpretation). In this case, for each $x \in X$ the set $\{Y \mid Y \subseteq X^* \wedge x \in Y\}$ is an ultrafilter of X^*.

DEFINITION 3.3 *A* canonical extension of B *is an algebra* B^σ*, which is a complete and atomic BA containing an isomorphic copy of* B *as a subalgebra, and which satisfies the following the properties:*

i) *Every atom of* B^σ *is the meet of elements of* B.

ii) *If* $A \subseteq B$ *such that* $\sum_{B^\sigma} A = 1$,
then there is a finite *set of* $A' \subseteq A$ *such that* $\sum_{B^\sigma} A' = 1$.

It is well known, that each BA has a canonical extension which is unique up to isomorphism. One such construction is given by Stone's representation theorem for Boolean algebras: Let B^σ be the powerset algebra of the set of ultrafilters X of B, and embed B into B^σ by $b \mapsto \{U \in X : b \in U\}$. If $A \leq B$, then A is called a *regular subalgebra of* B, if B is a canonical extension of A. The notion of canonical extension is equivalent to that of 'perfect' extension introduced in Jónsson and Tarski, 1951. Our notion of regular sub-algebra is also equivalent to that used in Jónsson and Tarski, 1951, which is different to that given in Koppelberg, 1989.

For more details and discussions we refer the reader to Jónsson, 1993, Jónsson, 1994, Jónsson, 1995, Jónsson and Tarski, 1951.

2.2 Boolean algebras with operators

The structure of a Boolean algebra may be further elaborated by the introduction of additional operators. These *Boolean algebras with operators* arose from the investigation of relation algebras, and were first studied in detail by Jónsson and Tarski, 1951; a survey can be found in Jónsson, 1993. Many useful structures have the form of such algebras.

DEFINITION 3.4 *Some useful concepts for describing Boolean algebras with operators are defined as follows:*

i) *A function $f : B^n \to B$ on a BA is called* additive *in its i-th argument if* $f(x_0, \ldots, x_i, \ldots x_n) + f(x_0, \ldots, x'_i, \ldots x_n) = f(x_0, \ldots, (x_i + x'_i), \ldots x_n)$, *for all $x_i \in B$.*

ii) *A function $f : B^n \to B$ on a BA is called an* operator, *if it is additive in each of its arguments.*

iii) *$f : B^n \to B$ is called* normal, *if it is additive and its value is 0 if any of its arguments is 0.*

iv) *A structure $\langle B, (f_i)_{i \in I} \rangle$ is called a* Boolean Algebra with Operators *(BAO), if B is a BA, and all f_i are operators.*

v) *If all f_i are furthermore normal, then we speak of a* normal BAO.

vi) *A collection of algebras defined by a given signature and a set of universally quantified equations is called an* equational class *(or* variety*).*

Examples of normal BAOs are modal algebras, relation algebras (both of which will be discussed below), and cylindric algebras, which provided an algebraisation of 1st-order logic. We invite the reader to consult the classic monographs by Henkin et al., 1971, Henkin et al., 1985 or the recent exposition by Andréka et al., 2001, which provides a comprehensive (and comprehensible) introduction to Tarski's algebraic logic.

The concept of canonical extensions of Definition 3.3 can be extended to BAOs:

DEFINITION 3.5 *Suppose that B is a BA, and f an n-ary normal operator on B. The* canonical extension *f^σ of f is defined by*

$$f^\sigma(x) = \sum \left\{ \prod \{ f(y) : p \leq y \in B^n \} : p \in \mathrm{At}(B^\sigma)^n \text{ and } p \leq x \right\} \tag{3.1}$$

for all $x \in (B^\sigma)^n$. If $\langle B, (f_i)_{i \in I} \rangle$ is a normal BAO, we call $\langle B^\sigma, (f_i^\sigma)_{i \in I} \rangle$ the canonical extension of $\langle B, (f_i)_{i \in I} \rangle$.

PROPOSITION 3.6 (JÓNSSON AND TARSKI, 1951) *The canonical extension of a normal BAO $\langle B, (f_i)_{i \in I} \rangle$ is a complete and atomic normal BAO containing $\langle B, (f_i)_{i \in I} \rangle$ as a subalgebra.*

This is not the place to dwell on the preservation properties of canonical extensions of normal BAOs, and we refer the reader to Jónsson, 1993 and de Rijke and Venema, 1995 for details.

As the connection of unary normal operators to operators of *modal logics* (which will be examined in detail later) is somewhat special, we make the following convention:

DEFINITION 3.7 *If f is a unary normal operator on the BA B, we call it a* modal operator *or* possibility operator*, and the structure $\langle B, f\rangle$ a* modal algebra*.*

Hence, modal algebras form an equational class (or variety) of algebras. That is the class of BAOs with one operator f that satisfy, in addition to the identities of Boolean algebra, the equations:

$$f(x+y) = f(x) + f(y), \tag{3.2}$$

$$f(0) = 0. \tag{3.3}$$

A special case of modal algebras (hence, of BAOs) are *closure algebras*:

DEFINITION 3.8 *A possibility operator f on B which also satisfies, for all $a \in B$,*

$$a + f(a) = f(a), \tag{3.4}$$

$$f(f(a)) = f(a). \tag{3.5}$$

is called a closure operator*, and, in this case, $\langle B, f\rangle$ is a* closure algebra*.*

Incidentally, one dimensional cylindric algebras are a special case of closure algebras (Henkin et al., 1971).

DEFINITION 3.9 *Functions whose duals are possibility operators are called* necessity *operators. Thus a necessity operator on B is a function $g : B \to B$, for which*

$$g(1) = 1, \qquad \textit{(Dually normal)} \tag{3.6}$$

$$g(a \cdot b) = g(a) \cdot g(b) \ \textit{for all } a, b \in B. \qquad \textit{(Multiplicative)} \tag{3.7}$$

DEFINITION 3.10 *A necessity operator g is called an* interior operator *if for all $a \in B$ it satisfies*

$$g(a) + a = a, \tag{3.8}$$

$$g(a) = g(g(a)). \tag{3.9}$$

If g is an interior operator on B, then the structure $\langle B, g\rangle$ is called an interior algebra*.*

Modal algebras can be viewed as an algebraic counterpart to the relational structures known as *(Kripke) frames*:

DEFINITION 3.11 *A* frame *is a pair $F = \langle U, R\rangle$, where R is a binary relation on U, called an* accessibility relation*.*

Every frame F has a corresponding algebra, called the *complex* algebra of F.

DEFINITION 3.12 *If $F = \langle U, R\rangle$ is a frame, then the* complex algebra *of F is the structure $F^* = \langle 2^U, \Diamond_R\rangle$, where $\Diamond_R : 2^U \to 2^U$ is defined by*

$$\Diamond_R(X) = \{y \in U : (\exists x \in X)xRy\}. \tag{3.10}$$

It is not hard to see that F^* is a complete and atomic modal algebra.

Conversely, we can construct a frame from a modal algebra:

DEFINITION 3.13 *If $\langle B, f\rangle$ is a modal algebra, let $R_f \in \mathrm{Rel}(\mathrm{At}(B))$ be defined by*

$$aR_fb \iff a \leq f(b). \tag{3.11}$$

The structure $\langle \mathrm{At}(B), R_f\rangle$ is called the canonical frame of $\langle B, f\rangle$, *and R_f its* canonical relation.

(We use the symbol $\iff$ to express meta-level equivalences and semantic definitions.)

We now have the following representation theorem:

PROPOSITION 3.14 (JÓNSSON AND TARSKI, 1951; JÓNSSON, 1993; JÓNSSON, 1994) *Let $\langle B, f\rangle$ be a complete and atomic modal algebra. Then, $\langle B, f\rangle$ is a regular subalgebra of the complex algebra of its canonical frame. Furthermore, if $\langle B, f\rangle$ is isomorphic to a regular subalgebra of a complex algebra of some frame $\langle U, R\rangle$, then $\langle U, R\rangle \cong \langle \mathrm{At}(B), R_f\rangle$.*

It is worth mentioning that all normal BAOs, not only the ones with unary operators, are representable as regular subalgebras of complex algebras of frames. Normality is essential, since non-normal BAOs do not admit such a representation (Madarász, 1998).

Correspondence theory investigates, which relational properties of R can be expressed by its canonical modal operator and its dual (see e.g. van Benthem, 1984). We have, for example:

$$R \text{ is reflexive} \iff (\forall X)[X \subseteq \Diamond_R(X)], \tag{3.12}$$

$$R \text{ is symmetric} \iff (\forall X)[\Diamond_R(-\Diamond_R(-X)) \subseteq X], \tag{3.13}$$

$$R \text{ is transitive} \iff (\forall X)[\Diamond_R\Diamond_R(X) \subseteq \Diamond_R(X)]. \tag{3.14}$$

These correspondences, as well as the following result, have appeared already in Jónsson and Tarski, 1951:

PROPOSITION 3.15 *A modal algebra is a closure algebra if and only if its canonical relation is reflexive and transitive.*

2.3 Binary relations and relation algebras

A *binary relation* R on a set U is a subset of $U \times U$, i.e. a set of ordered pairs $\langle x, y\rangle$ where $x, y \in U$. Instead of $\langle x, y\rangle \in R$, we shall often write xRy. The smallest binary relation on U is the empty relation $\emptyset$, and the largest relation is the universal relation $U \times U$, which we will normally abbreviate as U^2. The *identity relation* $\langle x, x\rangle : x \in U$ will be denoted by $1'$, and its complement, the *diversity relation*, by $0'$. *Domain* and *range* of R are defined by

$$\mathrm{dom}(R) = \{x \in U : (\exists y \in U)xRy\}, \tag{3.15}$$

$$\mathrm{ran}(R) = \{x \in U : (\exists y \in U)yRx\}. \tag{3.16}$$

Furthermore, we let $R(x) = \{y \in U : xRy\}$.

The set of all binary relations on U will be denoted by $\mathrm{Rel}(U)$. Clearly, $\mathrm{Rel}(U)$ is a Boolean algebra under the usual set operations:

$$-R = \{\langle x, y\rangle : \neg(xRz)\}, \tag{3.17}$$

$$R \cup S = \{\langle x, y\rangle : xRy \text{ or } xSy\}, \tag{3.18}$$

$$R \cap S = \{\langle x, y\rangle : xRy \text{ and } xSy\}. \tag{3.19}$$

If $R, S \in \mathrm{Rel}(U)$, the *composition of R and S* is defined as

$$R\ ;\ S = \{\langle x, y\rangle : (\exists z)[xRz \text{ and } zSy]\}. \tag{3.20}$$

The *converse of* R, written as $R^{\smile}$, is the set

$$R^{\smile} = \{\langle y, x\rangle : xRy\}. \tag{3.21}$$

A detailed analysis of relation algebras can be found in Henkin et al., 1971, and an overview in Jónsson, 1991. The following lemma sets out some decisive properties of composition and converse.

LEMMA 3.16

i) $;$ *is associative and distributes over arbitrary joins.*

ii) $1'\ ;\ R = R\ ;\ 1' = R$.

iii) $^{\smile}$ *is bijective, of order two, i.e.* $R^{\smile\smile} = R$, *and distributes over arbitrary joins.*

iv) $(R\ ;\ S)^{\smile} = S^{\smile}\ ;\ R^{\smile}$.

v) $(R\ ;\ S) \cap T = \emptyset \iff (R^{\smile}\ ;\ T) \cap S = \emptyset \iff (T\ ;\ S^{\smile}) \cap R = \emptyset$.

Note that any equation and any inequality between relations can be written as $T = U^2$ for some T. To do this, it is convenient to first to define the operation $R \otimes S$, which gives the *symmetric difference* of R and S:

$$R \otimes S = (R \cap -S) \cup (S \cap -R). \tag{3.22}$$

We then have the following equivalences:

$$R = S \iff -(R \otimes S) = U^2, \tag{3.23}$$

$$R \neq S \iff (U^2; ((R \otimes S); U^2)) = U^2. \tag{3.24}$$

Implicitly, we use here the concept of *discriminator algebras* which are a powerful instrument of algebraic logic (see Heinrich, 1978 and also Jónsson et al., 1991).

DEFINITION 3.17 *The* full algebra of binary relations on U *is the structure* $\langle \mathrm{Rel}(U), \cap, \cup, -, \emptyset, U^2, \ ; \ , \ ^{\smile}, 1' \rangle$.

A Boolean subalgebra of $\mathrm{Rel}(U)$ *which is closed under* $;$ *and* $^{\smile}$ *and contains* $1'$ *will be called an* algebra of binary relations *(BRA).*

Many properties of relations can be expressed by equations (or inclusions) among relations, for example,

$$\begin{aligned} R \text{ is reflexive} &\iff (\forall x) xRx, \\ &\iff 1' \subseteq R. \end{aligned} \tag{3.25}$$

$$\begin{aligned} R \text{ is symmetric} &\iff (\forall x, y)[xRy \leftrightarrow yRx], \\ &\iff R = R^{\smile}. \end{aligned} \tag{3.26}$$

$$\begin{aligned} R \text{ is transitive} &\iff (\forall x, y, z)[xRy \wedge yRz \rightarrow xRz], \\ &\iff R \ ; \ R \subseteq R. \end{aligned} \tag{3.27}$$

$$\begin{aligned} R \text{ is dense} &\iff (\forall x) x(-R)x \wedge \\ &\qquad (\forall x, y)[xRy \rightarrow (\exists z) xRzRy], \\ &\iff R \cap 1' = \emptyset \wedge R \subseteq R \ ; \ R, \\ &\iff R \cap (1' \cup -(R \ ; \ R)) = \emptyset. \end{aligned} \tag{3.28}$$

$$\begin{aligned} R \text{ is extensional} &\iff (\forall x, y)[R(x) = R(y) \rightarrow x = y], \\ &\iff [-(R \ ; \ -R^{\smile}) \cap -(R^{\smile} \ ; \ -R)] \subseteq 1'. \end{aligned} \tag{3.29}$$

Observe that all formulae above contain at most three variables. This is no accident, as the following result shows:

PROPOSITION 3.18 (TARSKI AND GIVANT, 1987; GIVANT, 2006)

i) *The 1st-order properties of binary relations on a set U that can be expressed by equations using the operators $\cap, \cup, -, \ ; \ , \ ^{\smile}$, and constants $\emptyset, U^2, 1'$ are exactly those which can be expressed with at most three distinct variables.*

ii) *If $\mathcal{R}$ is a collection of binary relations on U, then, the closure of $\mathcal{R} \cup \{1'\}$ under the operations $\cap, \cup, -, \ ; \ , \ ^{\smile}$ is the set of all binary relations on U*

which are definable in the (language of the) relational structure $\langle U, \mathcal{R}\rangle$ *by 1st-order formulae using at most three variables, two of which are free.*

If A is a complete and atomic BRA, in particular if A is finite, then the actions of the Boolean operators are uniquely determined by the atoms. To determine the structure of A, it is therefore enough to specify the composition and the converse operation.

When dealing with an atomic BRA, it is often convenient to specify the composition operation by means of *composition table* (CT), which, for any two atomic relations R_i, R_j, specifies the relation $R_i; R_j$ in terms of its constituent atomic relations. Formally, a composition table is a mapping $\mathrm{CT} : \mathrm{At}(A) \times \mathrm{At}(A) \to 2^{\mathrm{At}(A)}$ such that

$$T \in \mathrm{CT}(R, S) \iff T \subseteq (R\ ;\ S). \tag{3.30}$$

Since A is atomic, we have

$$R\ ;\ S = \bigcup \mathrm{CT}(R, S). \tag{3.31}$$

CT can be conveniently written as a quadratic array (the *composition table of* A), where rows and columns are labelled with the atoms of A, and the cells contain $\mathrm{CT}(R, S)$.

BRAs are one instance of the class of *relation algebras*, which may be seen of an abstraction of algebras of binary relations (Tarski, 1941):

DEFINITION 3.19 *A* relation algebra *(RA)*

$$\langle A, +, \cdot, -, 0, 1, \ ;\ , \breve{}, 1'\rangle$$

is a structure of type $\langle 2, 2, 1, 0, 0, 2, 1, 0\rangle$ *which satisfies:*

(RA0) $\langle A, +, \cdot, -, 0, 1\rangle$ *is a Boolean algebra.*

(RA1) $x\ ;\ (y\ ;\ z) = (x\ ;\ y)\ ;\ z.$

(RA2) $(x + y)\ ;\ z = (x\ ;\ z) + (y\ ;\ z).$

(RA3) $x\ ;\ 1' = x.$

(RA4) $x^{\breve{}\breve{}} = x.$

(RA5) $(x + y)^{\breve{}} = x^{\breve{}} + y^{\breve{}}.$

(RA6) $(x\ ;\ y)^{\breve{}} = y^{\breve{}}\ ;\ x^{\breve{}}.$

(RA7) $(x^{\breve{}}\ ;\ -(x\ ;\ y)) \leq -y.$

Observe that BRAs and RAs are BAOs: An RA is a BAO where the additional operators $\langle\ ;\ , \breve{}, 1'\rangle$ form an involuted monoid, and the connection between this

monoid and the Boolean operations is given by **RA7**. The somewhat cryptic character of **RA7**, can be made clearer by observing that, in the presence of the other axioms, it is equivalent to the cycle law

$$(x\ ;\ y)\cdot z = 0 \Longleftrightarrow (x^{\smile}\ ;\ z)\cdot y = 0 \Longleftrightarrow (z\ ;\ y^{\smile})\cdot x = 0. \tag{3.32}$$

Tarski announced in the late 1940s that set theory and number theory could be formulated in the calculus of relation algebras:

> 'It has even been shown that every statement from a given set of axioms can be reduced to the problem of whether an equation is identically satisfied in every relation algebra. One could thus say that, in principle, the whole of mathematical research can be carried out by studying identities in the arithmetic of relation algebras'. (Chin and Tarski, 1951)

We invite the reader to consult Tarski and Givant, 1987, and, for an overview Ahmed, 2004 or Givant, 2006. Another excellent reference for the theory of RAs is the book by Hirsch and Hodkinson, 2002.

2.4 Topological spaces

We assume familiarity with the basic ideas of topological spaces and only briefly recap some notational details. A topological space is a structure $\langle X, \tau\rangle$, where X is the base set, and τ the collection of *open* subsets of X. τ is closed under arbitrary unions and finite intersections. If τ is understood, we may use X to refer to the topological space. The elements of X will be denoted by lower case Greek letters (except τ), and its subsets by lower case Roman letters.

If $x \subseteq X$, its interior is denoted by $\mathrm{int}(x)$, and its closure by $\mathrm{cl}(x)$. Observe that int and cl are respectively an interior operator in the sense of (3.8)–(3.9), and a closure operator in the sense of (3.4)–(3.5). A subspace y of X is *dense in* X, if $\mathrm{cl}(y) = X$.

The *boundary* $\partial(x)$ of $x \subseteq X$ is the set $\mathrm{cl}(x) \setminus \mathrm{int}(x)$. If $\alpha \in X$, and $\alpha \in x \in \tau$, then x is called an *open neighbourhood of* α. X is called *connected* if it is not the union of two disjoint non-empty open sets.

2.4.1 Separation conditions. The general framework of topological spaces includes structures of many different kinds. In particular the open sets may be more or less densely distributed within the space. Significant, fundamental properties of this distribution can often be described in terms of the existence of certain disjoint sets separating arbitrary points and/or subsets of the space. Such properties are known as *separation conditions*.

Later, in Sec. 4, we will show how axiomatic properties of spaces described in terms of the contact relation correspond to separation conditions of their topological interpretations. To this end, the following conditions are especially relevant:

T_1**.** A topological space X is a called T_1 *space*, if for any two distinct points α, β, there are $x, y \in \tau$ such that $\alpha \in x$, $\beta \notin x$ and $\beta \in y$, $\alpha \notin y$. This is equivalent to the fact that each singleton set is closed.

T_2 **(Hausdorff).** X is called a T_2 or Hausdorff space, if any two distinct points have disjoint open neighbourhoods. It is well known that each T_2 space is a T_1 space, and that each regular T_1 space is a T_2 space.

Regular. A space X is *regular* if every point α and every closed set not containing α are respectively included in disjoint open sets.

It is well known (see e.g. Engelking, 1977) that X is regular, if and only if for each non-empty $u \in \tau$ and each $\alpha \in u$ there is some $v \in \tau$ such that $\alpha \in v \subseteq \mathrm{cl}(v) \subseteq u$.

Semi-Regular. A space is *semi-regular* if it has a basis of regular open sets—i.e. every open set is a union of regular open sets.

Regularity implies semi-regularity, but not vice versa.

Weakly Regular. We call X *weakly regular* if it is semi-regular and for each non-empty $u \in \tau$ there is some non-empty $v \in \tau$ such that $\mathrm{cl}(v) \subseteq u$. Weak regularity may be called a "pointless version" of regularity, and each regular space is weakly regular.

Completely Regular. X is called *completely regular*, if for every closed x and every point $\alpha \notin x$ there is a continuous function $f : X \to [0, 1]$ such that $f(\beta) = 0$ for all $\beta \in x$, and $f(\alpha) = 1$.

Normal. X is called *normal*, if any two disjoint closed sets can be separated by disjoint open sets.

Weakly Normal. X is called *weakly normal*, if any two disjoint regular closed sets can be separated by disjoint open sets. Weak normality was introduced as 'κ-normality' by Shchepin, 1972.

For any T_1 space, X, the following entailments hold:

X is *normal*

$\Longrightarrow X$ is *weakly normal*

$\Longrightarrow X$ is *completely regular*

$\Longrightarrow X$ is *regular*

$\Longrightarrow X$ is *weakly regular*

$\Longrightarrow X$ is *semi-regular*.

None of these implications can be reversed (see Düntsch and Winter, 2005 for examples).

A space which is T_1 and regular is called a T_3 space and a space which is T_1 and normal is a T_4 space. The various conditions T_i are successively stricter as i increases. Thus, $T_4 \Longrightarrow T_3 \Longrightarrow T_2 \Longrightarrow T_1$.

2.4.2 Regular sets and their algebras. A set $x \subseteq X$ is called *regular open* if $x = \mathrm{int}(\mathrm{cl}(x))$, and *regular closed* if $x = \mathrm{cl}(\mathrm{int}(x))$. Clearly, the set complement of a regular open set is regular closed and vice versa. The collection of regular open sets (regular closed sets) will be denoted by $\mathrm{RegOp}(X)$ ($\mathrm{RegCl}(X)$). It is well known (Koppelberg, 1989) that $\mathrm{RegOp}(X)$ and $\mathrm{RegCl}(X)$ can be made into (isomorphic) complete Boolean algebras by the operations

$$\begin{aligned} x + y &= \mathrm{int}(\mathrm{cl}(x \cup y)), & x + y &= x \cup y, \\ x \cdot y &= x \cap y, & x \cdot y &= \mathrm{cl}(\mathrm{int}(x \cap y)), \\ -x &= X \setminus \mathrm{cl}(x), & -x &= X \setminus \mathrm{int}(x), \\ \mathbf{0} &= \emptyset, & \mathbf{0} &= \emptyset, \\ \mathbf{1} &= X, & \mathbf{1} &= X. \end{aligned}$$

$\mathrm{RegOp}(X)$ does not fully determine the topology on X:

PROPOSITION 3.20 (VAKARELOV ET AL., 2002) *If y is a dense subspace of X, then* $\mathrm{RegOp}(X) \cong \mathrm{RegOp}(y)$.

If we only want to consider the regular closed sets (or regular open sets), it suffices to look at semi-regular spaces. Let us call the topology $r(\tau)$ on X which is generated by $\mathrm{RegOp}(\tau)$ the *semi-regularisation of* $\langle X, \tau \rangle$. Then the following useful fact holds:

PROPOSITION 3.21 (DÜNTSCH AND WINTER, 2005) *Suppose that* $\langle X, \tau \rangle$ *is a topological space. Then,* $\langle \mathrm{RegOp}(\tau) \rangle = \langle \mathrm{RegOp}(r(\tau)) \rangle$.

Proof (See also Engelking, 1977, p. 84.)
Let $a \subseteq X$. Then,

$$\begin{aligned} \mathrm{cl}_{r(\tau)}(a) = -\bigcup \Big\{ m \in \mathrm{RegOp}(\tau) : m \cap \mathrm{cl}_\tau(a) = \emptyset \Big\} \\ \supseteq -\bigcup \Big\{ m \in \tau : m \cap \mathrm{cl}_\tau(a) = \emptyset \Big\} = cl_\tau(a). \end{aligned}$$

Let $a \in \mathrm{RegOp}(\tau)$. Then,

$$\begin{aligned}\mathrm{int}_{r(\tau)}\,\mathrm{cl}_{r(\tau)}(a) = &\ \mathrm{int}_{r(\tau)}\left(-\bigcup\{m \in \mathrm{RegOp}(\tau) : m \cap \mathrm{cl}_\tau(a) = \emptyset\}\right),\\ = &\ \bigcup\{t \in \mathrm{RegOp}(\tau) : t \cap m = \emptyset \text{ for all } m \in \mathrm{RegOp}(\tau)\\ &\ \text{with } m \cap \mathrm{cl}_\tau(a) = \emptyset\},\\ = &\ a,\end{aligned}$$

since a and t are regular open, and thus, $t \subseteq \mathrm{cl}_\tau(a)$ implies $t \subseteq a$.

Conversely, let $a \in \mathrm{RegOp}(r(\tau))$. Then,

$$a = \mathrm{int}_{r(\tau)}\,\mathrm{cl}_{r(\tau)}(a) = \bigcup\{t \in \mathrm{RegOp}(\tau) : t \subseteq \mathrm{cl}_{r(\tau)}(a)\}.$$

Now, $\mathrm{int}_\tau\,\mathrm{cl}_\tau(a) \in \mathrm{RegOp}(\tau)$, and thus,

$$a \subseteq \mathrm{int}_\tau\,\mathrm{cl}_\tau(a) \subseteq \mathrm{int}_{r(\tau)}\,\mathrm{cl}_{r(\tau)}(a) = a.$$

If $a \in \mathrm{RegOp}(\tau)$, then, by the preceding consideration, $-_\tau a = -_{r(\tau)} a$, and thus, $\mathrm{cl}_\tau(a) = \mathrm{cl}_{r(\tau)}(a)$. This implies the claim. QED

2.4.3 Closure algebras and topologies. The study of topologies via the closure or interior operator is sometimes called *pointless topology* (see, for example, Johnstone, 1983). McKinsey and Tarski, 1944 had already shown that the closure algebras (as specified by Definition 3.8) give rise to the collection of closed sets of a topological space, by proving the following representation theorem:

PROPOSITION 3.22 (MCKINSEY AND TARSKI, 1944)

i) *If $\langle X, \tau\rangle$ is a topological space, then $\langle 2^X, \mathrm{cl}\rangle$ is a closure algebra, called the* closure algebra over $\langle X, \tau\rangle$.

ii) *If $\langle B, f\rangle$ is a closure algebra, then there is some T_1 space $\langle X, \tau\rangle$ such that $\langle B, f\rangle$ is a subalgebra of $\langle 2^X, \mathrm{cl}\rangle$.*

Dual statements holds for interior algebras and the topological int operator:

PROPOSITION 3.23

i) *If $\langle X, \tau\rangle$ is a topological space, then $\langle 2^X, \mathrm{int}\rangle$ is an interior algebra, called the* interior algebra over $\langle X, \tau\rangle$.

ii) *If $\langle B, g\rangle$ is an interior algebra, then there is some T_1 space $\langle X, \tau\rangle$ such that $\langle B, g\rangle$ is a subalgebra of $\langle 2^X, \mathrm{int}\rangle$.*

2.4.4 Heyting algebras and topologies. Another way of looking at these algebras is via a certain class of lattices: An algebra $\langle A, +, \cdot, \Rightarrow, \mathbf{0}, \mathbf{1}\rangle$ of type $\langle 2, 2, 2, 0, 0\rangle$ is called a *Heyting algebra* if $\langle A, +, \cdot, \mathbf{0}, \mathbf{1}\rangle$ is a bounded lattice, and $\Rightarrow$ is the operation of *relative complementation*, such that if $a, b \in A$, then

$$a \Rightarrow b \text{ is the largest } x \in A \text{ for which } a \cdot x \leq b. \tag{3.33}$$

In other words,

$$a \cdot x \leq b \text{ if and only if } x \leq a \Rightarrow b. \tag{3.34}$$

In Rasiowa and Sikorski, 1963 such algebras are called *pseudo-Boolean* algebras. If $a \in A$, then its *pseudo-complement* a^* is the element $a \Rightarrow \mathbf{0}$, i.e. a^* is the largest $x \in A$ for which $a \cdot x = 0$. Heyting algebras form an equational class—i.e. a collection of algebras defined by a set of universally quantified equations (for details see Rasiowa and Sikorski, 1963 or Balbes and Dwinger, 1974). Furthermore, if $\langle B, g\rangle$ is an interior algebra, then the collection $O(B)$ of its open sets forms a Heyting algebra with $\Rightarrow$ defined as

$$a \Rightarrow b = g(-a + b). \tag{3.35}$$

In view of Proposition 3.23 we now have the following representation theorem (McKinsey and Tarski, 1944):

PROPOSITION 3.24 *For each Heyting algebra A, there exists a T_1 space X such that A is isomorphic to a subalgebra of the Heyting algebra of open sets of X.*

2.5 Proximity spaces

Proximities were introduced by Efremovič, 1952. The intuitive meaning of a proximity Δ is that $x\Delta y$ holds for some $x, y \subseteq X$, when x *is close to* y in some sense. Their axiomatisation is very similar to that of Boolean contact algebra to be discussed in Sec. 4. The main source on proximity spaces is the monograph by Naimpally and Warrack, 1970.

From the point of view of this investigation proximity spaces play a very useful role. On the one hand, the proximity approach is close to that of point set topology, and mappings between proximity spaces and corresponding topological spaces are well established. On the other hand the formulation of proximity spaces is based on a binary relation between point sets, whose meaning can be correlated with the *contact* relation that is taken as a primitive in many axiomatic and algebraic approaches to representing topological relationships between regions (which will be considered further in Sec. 3 below). Hence, proximity

spaces provide a link between these axiomatic or algebraic formulations and point-set topological models of space.

Formally, a binary relation Δ on the powerset of a set X is called a *proximity*, if it satisfies the following axioms for $x, y, z \subseteq X$:

(P1) If $x \cap y \neq \emptyset$ then $x\Delta y$.

(P2) If $x\Delta y$ then $x, y \neq \emptyset$.

(P3) Δ is symmetric.

(P4) $x\Delta(y \cup z)$ if and only if $x\Delta y$ or $x\Delta z$.

(P5) If $x(-\Delta)y$ then $x(-\Delta)z$ and $y(-\Delta) - z$ for some $z \subseteq X$.

DEFINITION 3.25 The pair $\langle X, \Delta\rangle$ is called a *proximity space*.

Sometimes the term *proximity space* has been used to include structures that do not satisfy axiom **P5**. Those satisfying **P5** are sometimes called *Efremovič* proximity spaces (Efremovič, 1952).

DEFINITION 3.26 A proximity is called *separated* if it satisfies

$$(\mathbf{P}_{\text{sep}}) \qquad \{\alpha\}\Delta\{\beta\} \text{ implies } \alpha = \beta$$

Thus, in a separated proximity space, no two distinct singleton sets are related by the proximity relation.

2.5.1 The topology associated with a proximity space. Each proximity space determines a topology on X in the following way: we take the closure of any set x as the set of all points α, such that $\{\alpha\}$ is proximal to x:

$$\text{cl}(x) = \{\alpha \in X : \{\alpha\}\Delta x\}. \tag{3.36}$$

PROPOSITION 3.27 (NAIMPALLY AND WARRACK, 1970)

i) *The operation of* (3.36) *defines the closure operator of a topology* $\tau(\Delta)$ *on* X *(which is not necessarily* T_1*).*

ii) $\langle X, \tau(\Delta)\rangle$ *is a completely regular space.*

iii) *If* Δ *is separated, then* $\langle X, \tau(\Delta)\rangle$ *is a* T_1 *space.*

iv) $x\Delta y$ *if and only if* $\text{cl}(x)\Delta\,\text{cl}(y)$.

A proximity which is relevant to our investigation is the *standard proximity* on a normal T_1 space X (Naimpally and Warrack, 1970). For $x, y \subseteq X$, let

$$x\Delta y \iff \text{cl}(x) \cap \text{cl}(y) \neq \emptyset. \tag{3.37}$$

Observe that Δ is separated, since X is a T_1 space and thus singletons are closed.

3. Contact relations

The relation of 'contact' is fundamental to the spatial description of configurations of objects or regions. Contact relations have been studied in the context of qualitative approaches to geometry going back as far as the work of de Laguna, 1922, Nicod, 1924, Whitehead, 1978 and subsequently of Clarke, 1981. More recently the contact relation has been employed as a fundamental primitive in the field of Qualitative Spatial Reasoning (Randell et al., 1992b; Borgo et al., 1996; Cohn et al., 1997; Pratt and Schoop, 1998; Pratt and Schoop, 2000; Stell, 2000; Düntsch et al., 1999; Düntsch et al., 2001a; and see also Sec. 5 below). This has emerged as a significant sub-field of Knowledge Representation, which is itself a major strand of research in Artificial Intelligence. (In AI and Qualitative Spatial Reasoning, the contact relation is often called 'connection'. In the present work we use contact to avoid confusion with the slightly different notion of 'connection' employed in topology.)

The contact relation can be seen as a weaker and more fine-grained cousin of the 'overlap relation', which is straightforwardly defined in terms of the 'part of' relation, thus: $xOy \equiv_{def} \exists z[\, zPx \wedge zPy \,]$. The properties of the parthood relation were first formalised by Leśniewski, 1931, as the basic relation of his *Mereology* (see also Leśniewski, 1983; Leśniewski, 1992).

DEFINITION 3.28 *A* contact relation C *is a relation satisfying the following axioms:*

(C1)	$\forall x[\, xCx \,]$	(*Reflexivity*),
(C2)	$\forall xy[\, xCy \rightarrow yCx \,]$	(*Symmetry*),
(C3)	$\forall xy[\, \forall z[\, zCx \leftrightarrow zCy \,] \rightarrow x = y \,]$	(*Extensionality*).

These axioms correspond to axioms A0.1 and A0.2 given by Clarke, 1981 for the mereological part of his calculus of individuals.In theories whose domain includes an empty/null region, **C1** is normally weakened to $\forall x[x = \emptyset \vee xCx]$.

Our main interest will be contact relations which are defined on open or closed sets of a topological space. Primary examples are collections $\mathfrak{M}$ of non-empty regular closed (or regular open) sets of some topological space X.

If we identify regions with elements of $\mathrm{RegCl}(X)$, it is natural to define C as the relation that holds just in case two regions share at least one point:

$$xCy \Longleftrightarrow x \cap y \neq \emptyset, \tag{3.38}$$

whereas, if our domain of regions is $\mathrm{RegOp}(X)$, it is usual to define C as holding whenever the *closures* of two regions share a point:

$$xCy \iff \mathrm{cl}(x) \cap \mathrm{cl}(y) \neq \emptyset. \tag{3.39}$$

It is easy to see that these interpretations fulfil the contact relation axioms **C1–3**. In the sequel, they will be called *the standard contact relations on* $\mathrm{RegCl}(X)$ *and* $\mathrm{RegOp}(X)$ *respectively*.

It is often useful to consider contact relations over other, more specific domains. Take, for example, the set D of all closed discs in the Euclidean, and define C by (3.38). Then, C obviously is a contact relation on D.

When describing properties of the C relation, it is often convenient to refer to the set of all regions connected to a given region. Thus, we define

$$C(x) = \{y \mid xCy\}. \tag{3.40}$$

In terms of this notation, the extensionality axiom can be stated as:

$$\forall xy[(C(x) = C(y)) \leftrightarrow x = y]. \tag{3.41}$$

Many other useful relations can be defined in terms of contact (see Clarke, 1981; Randell et al., 1992b and 5.1 below). A particularly important definable relation is that which is normally interpreted as the *part* relation:

$$xPy \equiv_{def} \forall z[\, zCx \rightarrow zCy\,]. \tag{3.42}$$

This definition (by itself) ensures that P is reflexive and transitive—i.e. it is a *pre-order*. And if we assume the extensionality of C (i.e. **C3**) it can be proved that P is antisymmetric, so that it must be a *partial order*.

The C relation is a very expressive primitive for defining topological relationships between regions. In terms of C the following useful relations can be defined. These definitions have been used to define the relational vocabulary of the well-known *Region Connection Calculus*, which will be discussed further in Sec. 5 below.

(3.43)	$xPPy \equiv_{def} xPy \wedge \neg yPx$	x is a Proper Part of y
(3.44)	$xOy \equiv_{def} \exists z[zPx \wedge zPy]$	x Overlaps y
(3.45)	$xDRy \equiv_{def} \neg xOy$	x is DiscRete from y
(3.46)	$xDCy \equiv_{def} \neg xCy$	x is disconnected from y
(3.47)	$xECy \equiv_{def} xCy \wedge \neg xOy$	x is Externally Connected to y
(3.48)	$xPOy \equiv_{def} xOy \wedge \neg xPy \wedge \neg yPx$	x Partially Overlaps y

(3.49) $xEQy \equiv_{def} xPy \wedge yPx$ x is Equal to y

(3.50) $xTPPy \equiv_{def} xPPy \wedge \exists z[zECx \wedge zECy]$
x is a Tangential Proper Part of y

(3.51) $xNTPPy \equiv_{def} xPPy \wedge \neg\exists z[zECx \wedge zECy]$
x is a Non-Tangential Proper Part of y

(3.52) $xTPPI\,y \equiv_{def} yTPPx$
x is an Inverse Tangential Proper Part of y

(3.53) $xNTPPI\,y \equiv_{def} yNTPPx$
x is an Inverse Non-Tangential Proper Part of y

In the presence of the extensionality axiom, (3.49) is equivalent to simply $x = y$.

It should be noted that for the defined relations to have their intuitive meaning, one should not include in the domain a 'null' region that is not connected to any other region. If such a null region is present, it would be be part of every other region. Consequently xOy would hold for all x and y, and other relations defined in terms of O would also have counter-intuitive interpretations.

Under typical interpretations of the C relation (not including a null region in the domain), the relations defined by (3.46)–(3.53) form a jointly exhaustive and pairwise-disjoint partition of possible relations between any two spatial regions (i.e. every two regions satisfy exactly one of the relations). This set of eight relations introduced in Randell et al., 1992b is often known as RCC-8, and is widely referred to in the AI literature on Qualitative Spatial Reasoning (see also Sec. 5.1 below).

3.1 Contact relation algebras

If C is taken to be a relation in a relation algebra, the properties **C1–3** of the contact relation correspond to the following relation algebraic conditions:

(**CRA1**) $1' \leq C$, (*Reflexivity*),

(**CRA2**) $C = C^{\smile}$, (*Symmetry*),

(**CRA3**) $[-(C\ ;\ -C) \cap -(C\ ;\ -C)^{\smile}] \leq 1'$, (*Extensionality*).

DEFINITION 3.29 *A relation algebra generated from a single relation C satisfying conditions* **CRA1–3** *will be called a* contact relation algebra *(CRA).*

Contact Relation Algebras (CRAs) were introduced and studied in Düntsch et al., 1999, where many fundamental properties are demonstrated. CRAs provide a rich language within which many other useful topological relations can be defined. In the relation algebra setting, the part relation has the following definition:

(3.54) $$P \equiv_{def} -(C\ ;\ -C)$$

Many other relations are relationally definable from C. Indeed all the relations that were defined above using 1st-order logic can also be defined using the algebraic operators of relation algebra:

(3.55)	$PP =_{def} P \cap -1'$	proper part of
(3.56)	$O =_{def} P^{\smile} ; P$	overlap
(3.57)	$DR =_{def} -O$	discrete
(3.58)	$DC =_{def} -C$	disconnected
(3.59)	$EC =_{def} C \cap -O$	external contact
(3.60)	$PO =_{def} O \cap -(P \cup P^{\smile})$	partial overlap
(3.61)	$EQ =_{def} (P \cup P^{\smile}) \quad (= 1')$	equality
(3.62)	$TPP =_{def} PP \cap (EC ; EC)$	tangential proper part
(3.63)	$NTPP =_{def} PP \cap -TPP$	non-tangential proper part
(3.64)	$TPPI =_{def} TPP^{\smile}$	tangential proper part inv.
(3.65)	$NTPPI =_{def} NTPP^{\smile}$	non-tang'l proper part inv.

In view of Proposition 3.18, this comes as no surprise, since RAs capture exactly those 1st-order properties of C that can be expressed with up to three variables, and this is sufficient for all the definitions given above.

Depending on the base set, some of these relations might be empty or coincide. If, for example, B is a BA, and $xCy \iff x \cdot y \neq \mathbf{0}$, then C coincides with the overlap relation, and $EC = \emptyset$. A picture of some of these relations over the domain D of (non-empty) closed discs is given in Fig. 3.1.

It turns out that the relations

$$1', DC, PO, EC, TPP, TPP^{\smile}, NTPP, NTPP^{\smile} \tag{3.66}$$

are the atoms of the relation algebra $\mathcal{D}_c$ generated by C over D, henceforth called the *(closed) disc relations*. (The 'composition table' for the RCC-8 relations over the domain $\mathcal{D}_c$ will be given in Table 6.2 below.)

4. Boolean contact algebras

While the contact relations of Sec. 3 did not assume a particular algebraic structure on the base set, we will often be interested in cases where the set of regions has further structure; and, in particular, we will often want to consider the set of regions as having the structure of a Boolean algebra.

A 1st-order theory intended to model topological properties of regions, the *Region Connection Calculus* (RCC), has been introduced by Randell et al., 1992b in 1992, and has since gained popularity in the spatial reasoning community; we will examine the RCC more closely in Sec. 5. First, we will consider a more general class of structures:

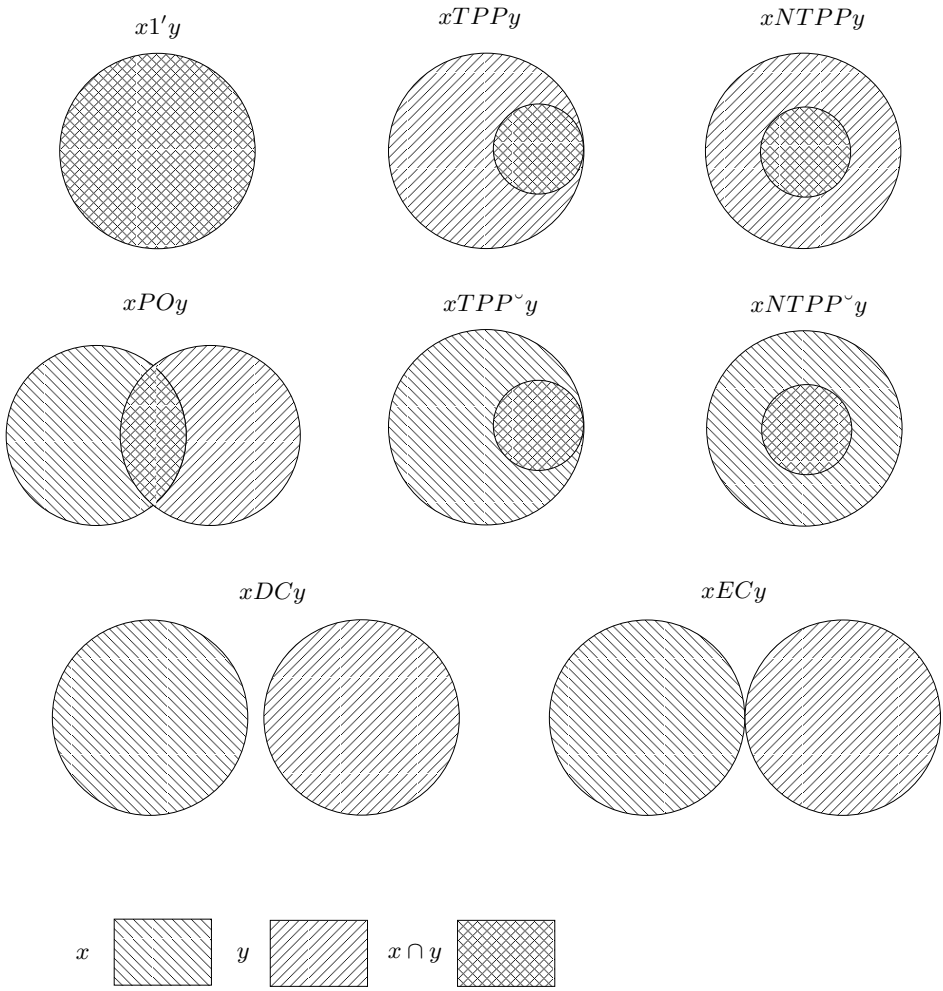


Figure 3.1. Topological relations on the domain of closed discs.

DEFINITION 3.30 *A* Boolean contact algebra *is a pair* $\langle B, C\rangle$*, such that* B *is a non-trivial (i.e.* $\mathbf{0} \neq \mathbf{1}$*) Boolean algebra, and* C *is a binary relation on* B^+*, called a* contact relation*, with the following properties:*

(BCA0)	aCb implies $a, b \neq \mathbf{0}$
(BCA1)	$a \neq \mathbf{0}$ implies aCa
(BCA2)	C is symmetric.

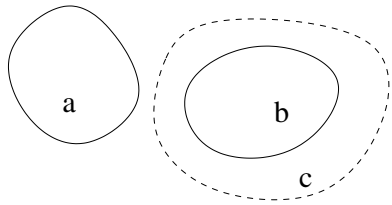


Figure 4.2. Illustration of the interpolation axiom.

(BCA3) (*Compatibility*) aCb and $b \leq c$ implies aCc

(BCA4) $aC(b+c)$ implies aCb or aCc

While axioms **BCA0–4** characterise the properties of Boolean contact algebras in general, we shall often be interested in BCAs that satisfy additional axioms. In particular, we shall be interested in the following axioms:

(BCA5)
(*Extensionality*) $C(a) \subseteq C(b)$ implies $a \leq b$

(BCA6)
(*Interpolation*) If $a(-C)b$ there is some c such that $a(-C)c$ and $-c(-C)b$

(BCA7)
(*Connection*) $a \notin \{\mathbf{0}, \mathbf{1}\}$ implies $aC - a$

A BCA which satisfies **BCA5** and **BCA7** will be called an *RCC algebra*, since these axioms are satisfied by the 1st-order Region Connection Calculus theory proposed by Randell et al., 1992b (which will be considered in further detail in Sec. 5 below).

Clearly, C is a contact relation in the sense of Sec. 3, and therefore, all relations specified by the definitional formulae (3.42)–(3.51) are at our disposal. It is easy to see that

(3.67) $\mathbf{BCA5} \iff P$ is the Boolean order,

(3.68) $\mathbf{BCA6} \iff \forall(x,y)(\exists z)[xNTPPz \wedge zNTPPy]$,

(3.69) $xOy \iff x \cdot y \neq \mathbf{0}$.

Simple structural properties include

PROPOSITION 3.31 *Let $\langle B, C\rangle$ be a BCA:*

i) (DÜNTSCH AND WINTER, 2004) *O is the smallest contact relation on B.*

ii) (DÜNTSCH AND WINTER, 2004) *B is a finite-cofinite algebra if and only if O is the only contact relation on B that satisfies* **BCA5**.

iii) (DÜNTSCH ET AL., 2001B) *If C satisfies* **BCA5** *and* **BCA7**, *then B is atomless.*

4.1 Interpretations of BCAs

As intended, the regions and relations of the BCA theory can be interpreted in terms of classical point-set topology. In fact, there are two dual interpretations that are equally reasonable.

Closed Interpretation:

- A *region* is identified with a regular closed set of points.
- Regions are *connected* if they share at least one point.
- Regions *overlap* if their *interiors* share at least one point.

Open Interpretation:

- A *region* is identified with a regular open set of points.
- Regions are *connected* if their *closures* share at least one point.
- Regions *overlap* if they share at least one point.

The axioms for C translate into topological properties as follows:

PROPOSITION 3.32 (DÜNTSCH AND WINTER, 2005) *Suppose that $\langle X, \tau\rangle$ is a topological space, and that C_τ is the standard contact relation on* $\mathrm{RegCl}(X)$.

i) *C_τ satisfies* **BCA0–4**.

ii) *C_τ satisfies* **BCA5** *if and only if X is weakly regular.*

iii) *C_τ satisfies* **BCA6** *if and only if X is weakly normal.*

iv) *C_τ satisfies* **BCA7** *if and only if X is connected.*

In fact the BCA axioms are also satisfied by dense subalgebras of $\mathrm{RegCl}(X)$. Hence, proposition 3.32 can be generalised:

PROPOSITION 3.33 *Suppose that $\langle X, \tau\rangle$ is a topological space, and C_τ is the standard contact relation on some* dense sub-algebra *of* $\mathrm{RegCl}(X)$*; then each of the clauses i–iv of proposition 3.32 are true for C_τ.*

The preceding propositions give us many examples of BCAs. We would like to mention a countable example of a BCA which is, in some sense, one dimensional; in particular, this algebra is not complete (i.e. does not contain infinite sums of its elements). Suppose that L is the ordered set of non-negative rational numbers enhanced by a greatest element ∞. Let B be the collection

of all finite unions of left-closed, right-open intervals of L, together with the empty set. It is well known (Koppelberg, 1989) that B is a Boolean subalgebra of 2^L, called the *interval algebra of* L, and that each $a \in B^+$ has a unique representation as

$$a = [x_0, y_0) \cup \ldots \cup [x_n, y_n), \tag{3.70}$$

where $x_0 \lneq y_0 \lneq x_1 \lneq y_1 \lneq \ldots \lneq x_n \lneq y_n$. The set $\{x_i : i \leq n\} \cup \{y_i : i \leq n\}$ is called the set of *relevant points of* a, denoted by $\mathrm{rel}(a)$. If we define C on B^+ by

$$aCb \iff (a \cap b) \cup (\mathrm{rel}(a) \cap \mathrm{rel}(b)) \neq \emptyset, \tag{3.71}$$

then $\langle B, C\rangle$ is a BCA which satisfies **BCA6** and **BCA7** (Düntsch and Winter, 2004). Other constructions of countable BCAs can be found in Li et al., 2005. In Sections 5 and 5.1 we will present BCAs arising from spatial theories.

We now exhibit some constructions that allow us to obtain new BCAs from old (these were described in Düntsch and Winter, 2004):

PROPOSITION 3.34 (ADDING AN ***ultra-contact***) *Given any atomless BCA $\langle B, C\rangle$ it is possible to augment the connection relation by picking any two ultrafilters F and G of the algebra and stipulating that $C(f, g)$ for any two regions f and g, where $f \in F$ and $g \in G$. In formal terms this means that $\langle B, C'\rangle$ is a BCA where*

$$C' = C \cup (F \times G) \cup (G \times F). \tag{3.72}$$

More generally, for a contact relation C, let $R_C = \{\langle F, G\rangle : F \times G \subseteq C\}$, and, for a reflexive and symmetric relation R on Ult(B), set $C_R = \bigcup\{F \times G : \langle F, G\rangle \in R\}$.

PROPOSITION 3.35

i) (DÜNTSCH AND VAKARELOV, 2006) *C_R satisfies* **BCA0–4**.

ii) (DÜNTSCH AND WINTER, 2006) *If R is a reflexive and symmetric relation on Ult(B) which is closed in the product topology of Ult$(B) \times$ Ult(B), then C_R satisfies* **BCA0–4**.

iii) (DÜNTSCH AND WINTER, 2006) *The collection of all relations on B that satisfy* **BCA0–4** *can be made into an atomistic complete co-Heyting algebra in which join is set union.*

iv) (DÜNTSCH AND WINTER, 2006) *The collection of all relations on B that satisfy* **BCA0–4** *and* **BCA6** *is isomorphic to the lattice of closed equivalence relations on Ult(B).*

PROPOSITION 3.36 (RESTRICTION AND EXTENSION) *If A is a dense subalgebra of B, then the restriction of C to A is a contact relation on A which satisfies* **BCA7** *if B does.*

If B is a dense subalgebra of A, then the relation C' defined on A by

$$aC'b \iff (\forall s, t \in B)[a \leq s \text{ and } b \leq t \Rightarrow sCt]$$

is a contact relation on A, and, if C satisfies **BCA7**, *so does C'. Furthermore, C' is the largest contact relation on A whose restriction to B is C.*

4.2 Representation theorems for BCAs

Theorems that characterise the class of models of a given axiomatic theory are know as *representation theorems*. In most cases, such theorems are sought after for one (or both) of the following reasons:

a) to find an axiomatisation for a given class of structures,

b) to show that a given axiom system is complete for an intended class of models.

Famous representation results include Cayley's theorem that every group is isomorphic to a group of permutations, and Stone's theorem which shows that each Boolean algebra is isomorphic to an algebra of sets. If an axiom system has models outside an intended class of models, the existence of such non-standard models shows that the system is incomplete with respect to that intended class. In the sequel, we will exhibit both positive and negative representation results for contact relations in topological spaces.

Apart from the earlier topological representation results of Roeper, 1997 and Mormann, 1998, which do not result in the standard topological contact, the first 'standard' representation result for a class of contact algebras was discovered by Vakarelov et al., 2001. It utilises the theory of proximity spaces which have been briefly described in Sec. 2. Subsequently, making use of similar techniques, topological representation results were obtained for BCAs (Düntsch and Winter, 2005).

4.2.1 Constructing a topology to represent a BCA. The proof of the representation result takes a form similar to that of Stone's theorem. The plan is to devise a way to use the elements of a BCA to construct entities that can be correlated with points in a topological or proximity space. However, instead of taking ultrafilters as the base set for the topology (as is done in Stone's theorem), a somewhat different construction is required to generate suitable sets of regions that can be identified with 'points' in a proximity space or topological model.

We begin with the following definition:

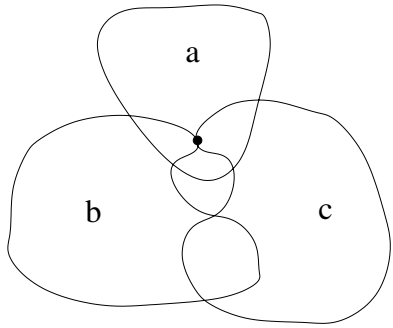


Figure 4.3. Illustration of why clans are not closed under intersection.

DEFINITION 3.37 *A non-empty subset* Γ *of* B *is called a* clan *if, for all* $x, y \in B$*, we have:*

CL1) *If* $x, y \in \Gamma$ *then* xCy*.*

CL2) *If* $x + y \in \Gamma$ *then* $x \in \Gamma$ *or* $y \in \Gamma$*.*

CL3) *If* $x \in \Gamma$ *and* $x \leq y$*, then* $y \in \Gamma$*.*

A clan can be regarded as a set of regions which share at least one point of mutual contact. The difference from a Boolean filter arises because regions may share a point of contact even though their intersection is empty. Moreover, as is illustrated in Fig. 4.3, even where regions do have a non-empty intersection, the regions may have a point of contact that is not in this intersection.

DEFINITION 3.38 *A clan* Γ *that is maximal (i.e. there is no clan* Γ' *such that* $\Gamma \subsetneq \Gamma'$*) will be called a* cluster*. The set of all clusters in* B *will be denoted by Clust*(B)*. Clearly, every clan is contained in some cluster.*

Since clusters will represent points in a topological space, each region will be associated with a set of clusters. Hence, to construct the topological representation of a BCA, we need to find a suitable mapping from the elements of the BCA to sets of clusters. Again the construction is similar to that used in the Stone theorem.

We define a mapping $h : B \to 2^{\text{Clust}(B)}$ by

$$h(a) = \{\Gamma \in \text{Clust}(B) : a \in \Gamma\}, \tag{3.73}$$

In Düntsch and Winter, 2005 it was shown that for any BCA with domain B, we can specify a topology $\langle \text{Clust}(B), \tau_B \rangle$, determined by h. This is done by taking $\{h(x) : x \in B\}$ as a basis for the closed sets of $\langle \text{Clust}(B), \tau_B \rangle$. In

other words the open sets τ_B are arbitrary unions of sets whose complements are in the range of h:

$$\tau_B = \left\{ \bigcup \{\text{Clust}(B) \backslash h(x) : x \in S\} : S \subseteq B \right\}$$

LEMMA 3.39 (DÜNTSCH AND WINTER, 2005) *The following properties of $\langle Clust(B), \tau_B \rangle$ hold.*

i) *The range of $h(x)$ for $x \in B$ is a dense subalgebra $\mathcal{A}_B$ of the regular closed algebra over $\langle Clust(B), \tau_B \rangle$.*

ii) *h preserves the Boolean structure of B in $\mathcal{A}_B$ (i.e. h is a Boolean homomorphism from B to $\mathcal{A}_B$).*

iii) *For all $a, b \in B$, aCb if and only if $h(a) \cap h(b) \neq \emptyset$.*

iv) *$\langle Clust(B), \tau_B \rangle$, is a weakly regular T_1 topology (which is not necessarily T_2),*

Together, these properties give us the following representation theorem:

PROPOSITION 3.40 *Each BCA $\langle B, C \rangle$ is isomorphic to a dense substructure of some regular closed algebra $\langle \mathrm{RegCl}(X), C_\tau \rangle$, where τ is a weakly regular T_1 topology, and C is the restriction of C_τ to B.*

Moreover, from propositions 3.32 and 3.33, we immediately have the following result which tells us that the correspondence is bijective:

PROPOSITION 3.41 *If $\langle X, \tau \rangle$ is a weakly regular T_1 space, and B is a dense subalgebra of $\mathrm{RegCl}(X)$ with C being the restriction of the standard contact on $\mathrm{RegCl}(X)$, then $\langle B, C \rangle$ is a BCA.*

As a consequence of this result we obtain

PROPOSITION 3.42 *The axioms of BCAs are complete with respect to the class of dense substructures of regular closed algebras of weakly regular T_1 spaces with standard contact.*

4.2.2 The extensionality axiom. The theorems stated in the last section concern BCAs satisfying the axioms **BCA0–5**—i.e. the general BCA theory together with the extensionality axiom. For certain purposes, in particular the modelling of discrete space, one may wish to remove the extensionality condition (Galton, 1999). The resulting very general BCAs will not be considered further here; however, representation results for general BCAs have been obtained by Düntsch and Vakarelov, 2006 and Dimov and Vakarelov, 2006.

4.2.3 The connection axiom. The connection axiom, **BCA7**, states that every region, except 0 and 1, is connected to its own complement:

$$a \notin \{\mathbf{0}, \mathbf{1}\} \text{ implies } aC - a.$$

Suppose **BCA7** is false for an RCA with domain B; then there are regions $a, b \in B$ such that $a, b \neq 0$, $a+b = 1$ and $a(-C)b$. Because the mapping h preserves Boolean identities and the contact relation, we must have regular closed regions $h(a)$ and $h(b)$ in $\langle \text{Clust}(B), \tau_B \rangle$ such that $h(a) + h(b) = \text{Clust}(B)$ and $h(a) \cap h(b) = \emptyset$. Therefore, $\langle \text{Clust}(B), \tau_B \rangle$ must be a disconnected space.

Conversely, it can be shown that if $\langle \text{Clust}(B), \tau_B \rangle$ is a connected topological space, then the BCA, B must satisfy the axiom **BCA7**. This is a bit more difficult to demonstrate, but is proved in Düntsch and Winter, 2005. (Since elements of the BCA form only a dense subalgebra of $\text{RegCl}(\langle \text{Clust}(B), \tau_B \rangle)$, we cannot necessarily associate an arbitrary regular closed subset of $\text{Clust}(B)$ with an element of the BCA from which the topology was constructed. This means that mapping topological constraints to BCA axioms often requires detailed analysis of the cluster construction.) Thus we have the following representation theorem for RCC algebras—i.e. BCAs satisfying axioms **BCA0–5** and **BCA7**:

PROPOSITION 3.43 *The axioms for RCC algebras are complete with respect to the class of dense substructures of regular closed algebras of* connected *weakly regular T_1 spaces with standard contact.*

4.2.4 Saturated clusters and the interpolation axiom. We now consider the effect of the interpolation axiom, **BCA6**. Recall that this stipulates that

$$\forall xy[x(-C)y \rightarrow (\exists z)[x(-C)z \wedge -z(-C)y]].$$

This is a separation condition ensuring that for any two disconnected regions in the algebra, we can find a third region disconnected from the first and including the second as a non-tangential part.

We shall later see that we can establish a correspondence between BCAs satisfying **BCA6** and proximity spaces. In order to do this we show that in the presence of this condition, the clusters derived from the algebra exhibit a property called *saturation*, which results leads to a natural 'well-behaved' structure of the set of clusters.

DEFINITION 3.44 *A clan is called* saturated *iff it satisfies the following condition:*

(P) If xCy for every $y \in \Gamma$, then $x \in \Gamma$.

If a clan Γ over B is saturated then for any $x \in B$ such that $x \notin \Gamma$ there is some $y \in \Gamma$ such that $\neg(xCy)$. Therefore, $\Gamma \cup \{x\}$ is not a clan. So Γ must be a maximal clan. Thus we have the following lemma (Düntsch and Winter, 2005):

LEMMA 3.45 *Every saturated clan is a cluster.*

In formulating the proximity representation theorem for BCAs, clusters corresponding to saturated clans will be will be taken as the points of a proximity space. Hence, we use the following terminology:

DEFINITION 3.46 *A cluster that is a saturated clan will be called a* proximity cluster*, or more briefly a p-cluster.*

Intuitively, each p-cluster can be interpreted as the set of all regions in the BCA that contain a particular point in a corresponding proximity space. However, for BCAs in general, not every cluster need be a p-cluster. The following example of a BCA which includes clusters that are not p-clusters is given in Düntsch and Winter, 2005.

Suppose that B is the interval algebra whose elements are finite unions of left closed, right open intervals, $[x, y)$ on the rational unit interval $[0, 1)$. Let $C(i, j)$ hold between elements just in case their closures share a point. Now, let a, b be points such that $0 < a < b < 1$. If F_a is the ultrafilter of B of all sets containing a, and F_b is the ultrafilter of B of all sets containing b, then, by Proposition 3.34, the relation $C' = C \cup (F_a \times F_b) \cup (F_b \times F_a)$ is a contact relation over B, and it can be shown that $\Gamma = F_a \cup F_b$ is a cluster. However, if $s \lneq a \lneq t \lneq b$, and $x = [s, a) \cup [t, b)$, then $\{x\} \times \Gamma \subseteq C'$ (i.e. x is connected to every member of Γ). But, neither $[s, a)$ nor $[t, b)$ is in Γ, so (because clans must satisfy $CL2$) we must have $x \notin \Gamma$.

Let us see how this anomaly arose. By adding the ultra-contact between points a and b, we stipulated that every region containing point a is in the C' contact relation with every region containing point b. But, in this algebra, contact also holds between regions that do not share a point, but whose closures share a point. However, the relation C' does not necessarily hold between intervals i, j such that the closure of i includes a and the closure of j includes b. This mismatch leads to a kind of discontinuity in the contact relation C' relative to the underlying topology of the interval algebra.

LEMMA 3.47 (DÜNTSCH AND WINTER, 2005) *If $\langle B, C\rangle$ satisfies* **BCA6***, then each cluster is a p-cluster.*

In order to see why **BCA6** ensures that all clusters are saturated we first give another useful lemma:

LEMMA 3.48 *For every region r and cluster Γ, $r \in \Gamma$ if and only if for any set of regions $S = \{r_1, \ldots, r_n\}$ such that $r \leq r_1 + \ldots + r_n$, there is a region $r_i \in S$ such that $(\forall x \in \Gamma)[r_i C x]$.*

Proof (sketch) Since clusters are maximal clans then, for any cluster Γ, if $\Gamma \cup \{r, \ldots\}$ satisfies *CL1–3* then $r \in \Gamma$. Moreover, to show that $r \in \Gamma$ it

suffices to show that $\Gamma \cup \{r\}$ satisfies *CL1-2*, since then $\Gamma \cup \{x : x \geq r\}$ clearly satisfies *CL1-3*. It can be shown that $\Gamma \cup \{r\}$ satisfies *CL1-2* just in case for every sum $(r_1 + \ldots + (r_n + r_{n+1})) = r$ there is some r_i such that $(\forall x \in \Gamma)[xCr_i]$, and this implies the lemma. QED

Using this, we can prove Lemma 3.47 as follows:

Proof Let $\langle B, C\rangle$ be a BCA satisfying **BCA6**. Let Γ be a cluster derived from this algebra and r a region such that $\forall x \in \Gamma[xCr]$. Suppose, in contradiction to Lemma 3.47 that $r \notin \Gamma$. Then, by Lemma 3.48, there are $r_1, \ldots, r_n$, with $r \leq r_1 + \ldots + r_n$, such that for each r_i there is some $x_i \in \Gamma$ with $r_i(-C)x_i$. Then by **BCA6** there are regions $s_1, \ldots, s_n$, such that $s_i(-C)x_i$ and $r_i(-C)-s_i$ (so each s_i contains r_i and separates it from x_i). Let $s = s_1 + \ldots + s_n$. Thus $r(-C) - s$. Now pick any region $y \in \Gamma$. Clearly $y = y_1 + \ldots + y_n + z$, where $y_i = y \cdot s_i$ and $z = y \cdot -s$. Because of *CL2* we must have either $z \in \Gamma$ or some $y_i \in \Gamma$. But since $y_i = y \cdot s_i$ and $s_i(-C)x_i$ we have $y_i(-C)x_i$; so $y_i \notin \Gamma$ (because of *CL1*). Thus we must have $z \in \Gamma$. However, since $z = y \cdot -s$, we have $z \leq -s$ and because $r(-C) - s$ we have $r(-C)z$. But this contradicts the premiss that $\forall x \in \Gamma[xCr]$. Hence the supposition that $r \notin \Gamma$ is impossible, so we have proved Lemma 3.47. QED

The converse of Lemma 3.47 is not true (Düntsch and Winter, 2006).

4.2.5 Representation in proximity spaces. As noted above, in Sec. 2, proximity spaces form a useful intermediary between topological spaces and axiomatic theories based on a contact relation, which has analogous properties to the proximity relation. Indeed contact can be regarded as a limiting case of proximity.

The theory of proximity spaces and their relation to topological spaces has been developed in detail in the seminal work of Naimpally and Warrack, 1970. This analysis makes heavy use of a notion of *cluster*, which is very similar to (and was the inspiration for) the cluster construct for BCAs given above. Because proximity spaces satisfy axiom **P5**, the clusters employed by Naimpally and Warrack are saturated. Hence, in the case of BCAs satisfying **BCA6**, many of the results of Naimpally and Warrack, 1970 can be used to demonstrate correspondences between BCAs, proximity spaces and topologies.

We first consider how we can derive a BCA from a proximity, space:

PROPOSITION 3.49 *Let $\langle X, \Delta\rangle$ be a proximity space with associated topology $\tau(\Delta)$, and $\mathrm{RegCl}(X)$ be the regular closed subsets of X according to the topology $\tau(\Delta)$. Then the algebra $\langle \mathrm{RegCl}(X), \Delta\rangle$ is a BCA called the* proximity connection algebra over $\langle X, \Delta\rangle$.

DEFINITION 3.50 $\langle \mathrm{RegCl}(X), \Delta \rangle$ *is called a* standard proximity connection algebra*, if*

$$x \Delta y \text{ iff } x \cap y \neq \emptyset, \text{for all } x, y \in \mathrm{RegCl}(X).$$

For our purposes, it suffices to consider only standard connection algebras. This is because of the following theorem:

PROPOSITION 3.51 (VAKARELOV ET AL., 2001) *Each proximity connection algebra is isomorphic to a standard proximity connection algebra.*

It follows immediately from the proximity axioms and the BCA axioms, that each standard proximity connection algebra is a BCA that satisfies the interpolation axiom **BCA6** (corresponding to the proximity axiom **P5**). We will demonstrate in the remainder of this section that, conversely, each BCA which satisfies **BCA6** can be embedded into a standard proximity connection algebra.

Given a BCA, $\langle B, C \rangle$, satisfying **BCA6**, our aim is to define a proximity on Clust(B). As with the representation in a topological space, the proximity space construction will again make use of *clusters* to represent points in the proximity space. Hence, each subset of the space will correspond to a set of clusters.

Since a cluster is interpreted as the set of regions containing a given point, the intersection of two clusters is the set of regions containing two points. More generally, given a set X of clusters representing a set of points, the common intersection $\bigcap X$ will be the set of all regions that contain all those points. Using this idea, we can for any BCA define a proximity relation between pairs of cluster sets, which corresponds to the contact relation of the BCA:

DEFINITION 3.52 *For any BCA,* $\langle B, C \rangle$ *that satisfies* **BCA6***, we define a proximity relation over Clust*(B) *in the following way:*
for each $X, Y \subseteq$ *Clust*(B)

$$(\Delta_{\text{rep}}) \qquad X \Delta_B Y \text{ iff } (\forall x, y \in B)[x \in \bigcap X \text{ and } y \in \bigcap Y \text{ imply } xCy].$$

Using this construction, the following lemma was proved by Vakarelov et al., 2002, based on an earlier result of Naimpally and Warrack, 1970:

LEMMA 3.53 $\langle$*Clust*$(B), \Delta_B \rangle$ *is a separated proximity space.*

Thus, the construction of clusters together with the definition of a proximity relation on sets of clusters enables us to derive a proximity space from any BCA satisfying **BCA6**. The structure $\langle \mathrm{Clust}(B), \Delta_B \rangle$ can be regarded as a canonical representation of the BCA B in terms of a (separated) proximity space.

As with the topological representation, the correspondence between the regions of the original BCA and subsets of the derived proximity space can be

specified by a function $h : B \to 2^{\mathrm{Clust}(B)}$, defined by $h(a) = \{\Gamma \in \mathrm{Clust}(B) : a \in \Gamma\}$. This mapping both preserves the Boolean structure of the BCA and also associates the contact relation of the BCA with the proximity relation of the proximity space.

We have now shown that each BCA $\langle B, C\rangle$ that satisfies **BCA6** is isomorphic to a standard proximity algebra over the proximity space $\langle \mathrm{Clust}(B), \Delta_B\rangle$. In Sec. 2we saw that each proximity space is associated with a corresponding topology and the properties of this topology were characterised by Proposition 3.36. This means that we can use the proximity space derived from a BCA to define a corresponding topological space. This gives us the following topological representation theorem for BCAs satisfying the interpolation axiom:

PROPOSITION 3.54 (VAKARELOV ET AL., 2001) *Each BCA which satisfies* **BCA6** *is isomorphic to a dense substructure of the regular closed algebra of a completely regular T_1 space X with standard contact as defined by* (3.38). *Furthermore, X is connected if and only if C satisfies* **BCA7**.

It should be noted that not every completely regular T_1 space is the representation space of a BCA which satisfies **BCA6**, since these spaces must be weakly normal (see Propositions 3.32–3), and there are spaces that are completely regular T_1, but not weakly normal (Shchepin, 1972). We have, however:

LEMMA 3.55 *The BCA axioms* **BCA0–6** *are complete with respect to the class of dense substructures of regular closed algebras of weakly normal T_1 spaces with standard contact.*

LEMMA 3.56 *The BCA axioms* **BCA0–7** *are complete with respect to the class of dense substructures of regular closed algebras of weakly normal* connected *T_1 spaces with standard contact.*

5. Other theories of topological relations

5.1 The Region Connection Calculus

The *Region Connection Calculus* (RCC) of Randell et al., 1992b is an axiomatisation of certain spatial concepts and relations in classical 1st-order predicate calculus. It has become widely known in the field of Qualitative Spatial Reasoning, a research area within the Knowledge Representation field of Artificial Intelligence. There is some variation in the full set of axioms used for the RCC theory. The formal apparatus of the original theory is complicated by the use of the many-sorted logic LLAMA (Cohn, 1987) and the use of a non-standard definite description operator ($\iota x[\varphi(x)]$). This makes it difficult to make a direct comparison with the algebraically based theories presented in the current paper.

The RCC theory is based on a primitive relation C, which is in this context normally called the *connection* relation. This is axiomatised to be reflexive (**C1**) and symmetric (**C2**). The extensionality axiom (**C3**) is not given in the original RCC theory (Randell et al., 1992b) and does not strictly follow from the other axioms (see Bennett, 1997; Stell, 2000). However, the theory does contain definition 3.49 for the EQ relation; and, if (as seems to have been assumed in some subsequent development of RCC) this is taken as coinciding with logical equality, then **C3** also holds. With this assumption, we have a contact relation in the sense defined in Sec. 3.

The RCC theory introduces further relations by means of the definitions (3.42)–(3.53) given above (Sec. 3), which include of course the RCC-8 relation set. The following axiom is given stipulating that every region has a non-tangential proper part:

(RCC1) $$\forall x \exists y [yNTPPx].$$

However, as shown in Düntsch et al., 2001b, this follows from the other axioms, if we assume the extensionality axiom **C3**.

RCC also incorporates a constant denoting the *universal* region, a *sum* function and partial functions giving the *product* of any two overlapping regions and the *complement* of every region except the universe. With slight modification to the original to replace the partial product and complement functions with relations, these are defined as follows:

(RCCD1) $x = \mathcal{U} \equiv_{def} \forall y[xCy]$,

(RCCD2) $x = y + z \equiv_{def} \forall w[wCx \leftrightarrow [wCy \vee wCz]]$,

(RCCD3) $\mathsf{Prod}(x, y, z) \equiv_{def} \forall u[uCz \leftrightarrow \exists v[vPx \wedge vPy \wedge uCv]]$,

(RCCD4) $\mathsf{Compl}(x, y) \equiv_{def} \forall z[(zCy \leftrightarrow \neg zNTPPx) \wedge (zOy \leftrightarrow \neg zPx)]$.

It should be noted that within the original RCC theory there is no such thing as a *null* (or empty) region. Thus there is no product of discrete regions or complement of the universal region. This means we do not have a full Boolean algebra of regions; but, in order that appropriate regions exist to fulfil the requirements of the *quasi-Boolean* structure suggested by the above definitions, the basic RCC theory should be supplemented with the following existential axioms:

(RCC2) $$\forall xy[xOy \rightarrow \exists z[\mathsf{Prod}(x, y, z)]],$$

(RCC3) $$\forall x[\neg(x = \mathcal{U}) \leftrightarrow \exists y[\mathsf{Compl}(x, y)]].$$

The many-sorted formalisation of RCC and the choice to exclude the 'null region' from the domain of regions was motivated partly by a desire to accord

with 'commonsense' notions of spatial reality (influenced by e.g. Hayes, 1985) and partly by wanting to improve the effectiveness of automated reasoning using the calculus. However, from the point of view of establishing properties of the formal system, it has been found that the lack of a null region is problematic since it considerably complicates the comparison with standard mathematical structures such as Boolean algebras. Hence, subsequent investigations (e.g. Stell, 2000; Düntsch et al., 1999) have often modified the original theory by introducing a null region so that the theory can be built upon a domain that has the basic Boolean algebra structure.

Once the null region has been added, it is clear that the models of the revised RCC theory will be BCAs (as defined by Definition 3.30). Moreover, given the RCC axioms, it can be proved that every region is connected to its own complement:

$$\forall xy[\mathsf{Compl}(x,y) \rightarrow xCy]. \tag{3.74}$$

This corresponds to the BCA property **BCA7**. In fact any model of the RCC axioms modified to include the null region correspond to a BCA satisfying this property:

LEMMA 3.57 *An* RCC model *is an RCC algebra, i.e. a BCA* $\langle B, C\rangle$ *which satisfies* **BCA7**.

This correspondence enables us to use connected BCAs as an algebraic counterpart to the 1st-order RCC axioms. An another algebraic analysis of the RCC theory, employing a somewhat weaker axiomatisation, is given in Stell, 2000.

5.2 The 4- and 9-intersection representations

The 4 and 9 Intersection representations were originally described by Egenhofer and his students (Egenhofer and Franzosa, 1991; Egenhofer and Herring, 1991) as a means of representing relationships between geographic regions. The approach is based on the idea of interpreting regions as point sets and characterising binary spatial relations in terms of topological constraints on these sets. The originators suggest that the representation should be applied to Jordan curve bounded regions in the plane (i.e. regions that are homeomorphic to closed discs); however, there is no reason why it could not be applied more generally to regular closed subsets of a topological space.

In the 4-intersection representation the idea is to consider the intersection of the boundary and interior of one region with the boundary and interior of another. Thus, for regions A and B, we consider $\partial(A) \cap \partial(B)$, $\partial(A) \cap \mathrm{int}(B)$, $\mathrm{int}(A) \cap \partial(B)$, $\mathrm{int}(A) \cap \mathrm{int}(B)$, and we determine whether or not these intersections are empty (denoted $\emptyset$) or non-empty (denoted $\neg\emptyset$). The determined values are naturally represented by a 2x2 matrix, as shown in Table 5.1.

$\cap$	$\partial(B)$	$\mathrm{int}(B)$	$\cap$	$\partial(B)$	$\mathrm{int}(B)$
$\partial(A)$	$\neg\emptyset$	$\emptyset$	$\partial(A)$	$\neg\emptyset$	$\neg\emptyset$
$\mathrm{int}(A)$	$\emptyset$	$\neg\emptyset$	$\mathrm{int}(A)$	$\neg\emptyset$	$\neg\emptyset$
	$\mathbf{1}'$			$\mathbf{PO}$	
$\cap$	$\partial(B)$	$\mathrm{int}(B)$	$\cap$	$\partial(B)$	$\mathrm{int}(B)$
$\partial(A)$	$\emptyset$	$\neg\emptyset$	$\partial(A)$	$\neg\emptyset$	$\neg\emptyset$
$\mathrm{int}(A)$	$\emptyset$	$\neg\emptyset$	$\mathrm{int}(A)$	$\emptyset$	$\neg\emptyset$
	$\mathbf{NTPP}$			$\mathbf{TPP}$	
$\cap$	$\partial(B)$	$\mathrm{int}(B)$	$\cap$	$\partial(B)$	$\mathrm{int}(B)$
$\partial(A)$	$\emptyset$	$\emptyset$	$\partial(A)$	$\neg\emptyset$	$\emptyset$
$\mathrm{int}(A)$	$\emptyset$	$\emptyset$	$\mathrm{int}(A)$	$\emptyset$	$\emptyset$
	DC			EC	
$\cap$	$\partial(B)$	$\mathrm{int}(B)$	$\cap$	$\partial(B)$	$\mathrm{int}(B)$
$\partial(A)$	$\emptyset$	$\emptyset$	$\partial(A)$	$\neg\emptyset$	$\emptyset$
$\mathrm{int}(A)$	$\neg\emptyset$	$\neg\emptyset$	$\mathrm{int}(A)$	$\neg\emptyset$	$\neg\emptyset$
	$\mathbf{NTPP}^{\smile}$			$\mathbf{TPP}^{\smile}$	

Table 5.1. Topological relations definable using the 4-intersection representation.

By reference to Fig. 3.1 (in Sec. 3 above), it is easy to see that the base relations of the closed circle algebra can be described by this *4-intersection model* (Egenhofer and Franzosa, 1991).

An approach which extends the 4-intersection model also takes into account the complement of the sets in question, and can be described by the following matrix:

$$\begin{pmatrix} \mathrm{int}(x) \cap \mathrm{int}(y) & \mathrm{int}(x) \cap \partial(y) & \mathrm{int}(x) \cap -y \\ \partial(x) \cap \mathrm{int}(y) & \partial(x) \cap \partial(y) & \partial(x) \cap -y \\ -x \cap -\mathrm{int}(y) & -x \cap \partial(y) & -x \cap -y \end{pmatrix}$$

While the 4-intersection model described the topological invariant relations among closed Jordan curves, the 9-intersection model is able to describe such relations for sets, which have arbitrary shaped interiors, including lines and points. Details can be found in Egenhofer and Herring, 1991.

Thus we see that the 9-intersection representation, based on a point-set interpretation of regions characterises exactly the same set of basic binary relations as the axiomatic RCC theory. This of course is not surprising given the correspondences between axiomatic algebras and topological spaces characterised by the representation theorems given in Sec. 4.

6. Reasoning about topological relations

The foregoing sections have defined a rich array of formal frameworks for representing topological relationships between regions. We now look as ways in which these representations can be employed to make inferences about topological configurations of regions.

6.1 Compositional reasoning

Compositional inference may be described in general terms as a deduction, from two relational facts of the forms aRb and bSc, of a relational fact of the form aTc, involving only a and c. Such inferences may be useful in their own right or may be employed as part of a larger inference mechanism, such as a consistency checking procedure for sets of relational facts. In either case, one will normally want to deduce the strongest relation aTc that is entailed by $aRb \wedge bSc$ and which is expressible in whatever formalism is being employed.

In 1st-order logic we can express the strongest fact derivable from $aRb \wedge bSc$ by the formula $a(R;S)b$, where the ; operator is defined by:

$$x(R;S)y \equiv_{def} \exists z[xRz \wedge zSy]. \tag{3.75}$$

Hence, the meaning of ';' coincides with that of composition in Binary Relation Algebras (as defined in Sec. 2). This may be called the *strong* composition operator. It is also often called the *extensional* composition, because, if we know that x and y stand in a relation equivalent to $R;S$, we can infer the existence of an entity z, such that xRz and zSy.

As a means of practical reasoning, inferring strong compositions in an expressive language such as 1st-order logic may not be very effective as the formulae generated will in general be more complex than the initial formulae and no more informative. However, if it is found that for a certain set of relations, every formula derived by compositional inference is equivalent to some relatively simple formula (preferably a single relation of the language or perhaps a disjunction of relations) then compositional inference may be a very powerful tool.

In the case of the Allen calculus based on 13 basic temporal interval relations (Allen, 1983), it turns out that (under a very natural interpretation in terms of intervals on the rational line) the extensional compositions of any pair of base relations correspond to some disjunction of the basic 13 relations. Hence, composition can be applied without generating more complex relations.

It has been found that in cases where the strong, extensional composition cannot be simply expressed, it is useful to generalise the notion of composition to allow weaker inferences. In particular the following notion is often used:

DEFINITION 3.58 *Given a theory* Θ *whose vocabulary includes a set* Rels *of relations, the* weak composition*,* $WComp(R,S)$*, where* $R, S \in$ Rels *is defined to be the disjunction of all relations* $T_i \in$ Rels*, such that there exist individual constants a, b, c, where the formula* $R(a,b) \wedge S(b,c) \wedge T_i(a,c)$ *is consistent with* Θ*.*

This means that if $WComp(R,S) = T_1 \cup \ldots \cup T_n$ then

$$\Theta \models \forall x \forall y \forall z[xRy \wedge ySz \rightarrow (xT_1z \vee \ldots \vee xT_nz)]. \tag{3.76}$$

and, furthermore, $T_1 \cup \ldots \cup T_n$ is the smallest subset of Rels for which such a formula is provable. Moreover, it is easy to show that $R;S$ is always a sub-relation (or equivalent to) $WComp(R,S)$.

Given this definition, inferences of the following form will always be valid:

$$\frac{R(a,b) \wedge S(b,c) \wedge T(a,c)}{(WComp(R,S) \cap T)(a,c)} \text{[WComp]} \tag{3.77}$$

For a finite set of relations, $WComp(R,S)$ can be pre-computed for every pair of relations and stored in a matrix known as a *composition table*. This provides a simple mechanism for computing compositional inferences by looking up compositions in the table. The typical mode by which this kind of compositional reasoning is executed is to repeatedly infer compositional inferences using table look-up until either an inconsistency is detected or no new inferences can be made. Since their introduction by Allen, 1983, composition tables have received considerable attention from researchers in AI and related disciplines (Vilain and Kautz, 1986; Egenhofer, 1991; Freksa, 1992; Randell et al., 1992a; Röhrig, 1994; Cohn et al., 1994; Düntsch et al., 1999; Düntsch et al., 2001b). (In fact Allen called his table a 'transitivity table', but 'composition table' is arguably more appropriate and it seems that this is becoming the standard term.)

Table 6.2 is usually called *The Composition Table* for the RCC-8 relations. (The identity relation is omitted from the table since it is clear how composition with $1'$ works.) In general, the table gives the *weak* composition of the RCC-8 relations. This is because over many domains (e.g. over regular closed sets of an arbitrary topological space) the algebra is not atomic, so that compositional combinations of the RCC-8 relations generate an infinite set of different relations, many of which are not expressible as disjunctions of the RCC-8 generating set.

Nevertheless, there are some more restricted interpretations under which the RCC-8 relations are indeed the atoms of an RA (so that every relation in the algebra is a disjunction of these relations). One simple example is the case where the domain of regions is the set $\mathcal{D}_c$ of closed circles in the plane (with the usual interpretation of the relations) as depicted in Fig. 3.1. In the case where the domain is taken as the Jordan curve bounded regions of the plane, the table also corresponds to extensional composition (Li and Ying, 2003b; Li and Ying, 2003a).

It needs to be mentioned, that the RCC-8 relations are never the atoms of an RA generated by C over an RCC algebra (Düntsch, 2005), thus the composition table is not extensional for such algebras. For instance the composition table shows that $WComp(EC,EC) = 1' \cup DC \cup EC \cup PO \cup TPP \cup TPP^{\smile}$. If this were extensional, it would mean that for any regions a, b, such that $aECb$

we could find a third region c such that $aECc$ and $cECb$. However, suppose $a = -b$, then there can be no region in the relation EC to both a and b.

;			C				
	DR		O				
				PP		$PP^{\smile}$	
	DC	EC	PO	TPP	NTPP	$TPP^{\smile}$	$NTPP^{\smile}$
DC	1	DR, PO, PP	DR, PO, PP	DR, PO, PP	DR, PO, PP	DC	DC
EC	DR, PO, $PP^{\smile}$	1', DR, PO, TPP $TPP^{\smile}$	DR, PO, PP	EC, PO, PP	PO, PP	DR	DC
PO	DR, PO, $PP^{\smile}$	DR, PO, $PP^{\smile}$	1	PO, PP	PO, PP	DR, PO, $PP^{\smile}$	DR, PO, $PP^{\smile}$
TPP	DC	DR	DR, PO, PP	PP	NTPP	1', DR, PO, TPP, $TPP^{\smile}$	DR, PO, $PP^{\smile}$
NTPP	DC	DC	DR, PO, PP	NTPP	NTPP	DR, PO, PP	1
$TPP^{\smile}$	DR, PO, $PP^{\smile}$	EC, PO, $PP^{\smile}$	PO, $PP^{\smile}$	1', PO, TPP, $TPP^{\smile}$	PO, PP	$PP^{\smile}$	$NTPP^{\smile}$
$NTPP^{\smile}$	DR, PO, $PP^{\smile}$	PO, $PP^{\smile}$	PO, $PP^{\smile}$	PO, $PP^{\smile}$	O	$NTPP^{\smile}$	$NTPP^{\smile}$

Table 6.2. The composition table of $\mathcal{D}_c$.

6.2 Equational reasoning

In Sec. 2 we looked at the algebraic characterisation of topological spaces in terms of Closure Algebras and their complementary Interior Algebras. Since these algebras can be defined by purely equational axioms, this representation suggests that it should be possible to use some form of equational inference to reason about topological relationships among regions.

In general (according to Proposition 3.22) the elements of a closure algebra can correspond to arbitrary subsets of a topological space. However, in order that the domain of regions be compatible with the topological interpretations of region-based axiomatic theories (such as the BCAs discussed in Sec. 4.1 and 4.2), we will often want to identify and reason about either regular open or regular closed sets. In the first case one should assert an equation $x = \mathrm{int}(\mathrm{cl}(x))$ for each region variable x; and in the second case one should assert

$x = \mathrm{cl}(\mathrm{int}(x))$. In either case we can define a large vocabulary of relations in terms of equations of an interior/closure algebra.

If we are dealing with regions corresponding to regular closed sets then the following definitions of binary topological relations can be given:

$$xDCy \iff -(x \cdot y) = \mathbf{1}, \tag{3.78}$$
$$xDRy \iff -(\mathrm{int}(x) \cdot \mathrm{int}(y)) = \mathbf{1}, \tag{3.79}$$
$$xPy \iff -x + y = \mathbf{1}, \tag{3.80}$$
$$xP\breve{}y \iff x + -y = \mathbf{1}, \tag{3.81}$$
$$xNTPy \iff -x + \mathrm{int}(y) = \mathbf{1}, \tag{3.82}$$
$$xNTP\breve{}y \iff \mathrm{int}(x) + -y = \mathbf{1}, \tag{3.83}$$
$$xEQy \iff x = y. \tag{3.84}$$

(The extension of NTP coincides with $NTPP$, except that $1NTP1$ is true.)

But C itself (as well as many other relations, including O) cannot be defined by an interior algebraic equation. This follows from the general observation that purely equational constraints are always consistent with any purely equational theory (there must always be at least a trivial one-element model, in which all constants denote the same individual). Thus if the negation of some constraint can be expressed as an equation, then the constraint itself cannot be equationally expressible (otherwise that constraint would be consistent with its own negation).

So to define C (and O) we need to employ disequalities:

$$xCy \iff -(x \cdot y) \neq \mathbf{1}, \tag{3.85}$$
$$xOy \iff -(\mathrm{int}(x) \cdot \mathrm{int}(y)) \neq \mathbf{1}. \tag{3.86}$$

Moreover, all the RCC-8 relations can be defined by some combination of equations given in (3.79)–(3.84) and negations of these equations. Those not already specified, can be defined as follows:

$$xECy \iff (-(x \cdot y) = \mathbf{1}) \wedge (\mathrm{int}(-x) + \mathrm{int}(-y) \neq \mathbf{1}), \tag{3.87}$$
$$xPOy \iff (-(x \cdot y) \neq \mathbf{1}) \wedge (-x + y \neq \mathbf{1}) \wedge (x + -y \neq \mathbf{1}), \tag{3.88}$$
$$xTPPy \iff (-x + y = \mathbf{1}) \wedge (x \neq y) \wedge (\mathrm{int}(-x) + y \neq \mathbf{1}), \tag{3.89}$$
$$xTPP\breve{}y \iff (x + -y = \mathbf{1}) \wedge (x \neq y) \wedge (x + \mathrm{int}(-y) \neq \mathbf{1}). \tag{3.90}$$

For many applications we will also want to specify that certain regions are non-empty. This is easily done using the disequality $-x \neq 1$. Various other useful binary RCC relations are expressible by means of interior algebra equations. For example, $EQ(x + y, \mathbf{1})$ can be expressed by $X \cup Y = \mathbf{1}$.

The problem of reasoning with topological relations can thus be reduced to one of reasoning with algebraic equations and disequalities; and this in turn can be reduced to the problem of testing consistency of sets of equations and disequalities. Moreover, the following Lemma tells us that the consistency of such sets can be determined as long as we have a means of computing whether a given equation follows from a set of equations:

LEMMA 3.59 *A set of algebraic equalities and disequalities,* $\{x_1 = y_1, \ldots x_m = y_m, z_1 \neq w_1, \ldots x_n \neq y_n\}$*, is inconsistent just in case* $x_1 = y_1, \ldots x_m = y_m \models z_i = w_i$*, for some* z_i*, such that* $1 \leq i \leq n$*.*

Clearly this kind of approach could be applied to any of the other purely equationally defined algebras defined above (in Sec. 2).

Though equational reasoning has long been a major topic in mathematics and computer science and many general techniques are known, it seems that there has been little research directed specifically at equational reasoning in this kind of spatial algebra. However, an indirect way of implementing such reasoning is by means of an encoding into modal logic, described in the next section.

6.3 Encoding in propositional modal logics

A propositional modal logic augments the classical propositional logic with one or more unary connectives. We assume familiarity with the basics of these formalisms. Full details can be found in many texts, such as Hughes and Cresswell, 1968; Blackburn et al., 2001.

We first consider normal modal logics with a single modality. As usual, the modal necessity operator will be denoted by $\Box$, and its dual possibility operator by $\Diamond$ (where $\Diamond p \leftrightarrow \neg \Box \neg p$). Let $\mathfrak{F}$ be the set of all (well-formed) propositional modal formulae (defined in the usual way).

DEFINITION 3.60 *A* normal propositional modal logic $\mathcal{ML}$ *is identified with the set of its theorems. More specifically,* $\mathcal{ML}$ *is a subset of* $\mathfrak{F}$*, satisfying the following conditions:*

(ML1)		*All classical tautologies are in* $\mathcal{ML}$*.*
(ML2)		*If* $p \in \mathcal{ML}$ *and* $(p \rightarrow q) \in \mathcal{ML}$*, then* $q \in \mathcal{ML}$*.*
(ML3)		$\mathcal{ML}$ *is closed under substitution.*
(ML4)	*(Defn. of* $\Diamond$*)*	$(\Box p \leftrightarrow \neg \Diamond \neg p) \in \mathcal{ML}$*.*
(ML5)	*(Extensionality)*	*If* $(p \leftrightarrow q) \in \mathcal{ML}$ *then* $(\Diamond p \leftrightarrow \Diamond q) \in \mathcal{ML}$*.*
(ML6)		$\Diamond(p \wedge \neg p) \leftrightarrow (p \wedge \neg p)$*.*
(ML7)		$(\Diamond(p \vee q) \leftrightarrow (\Diamond p \vee \Diamond q)) \in \mathcal{ML}$*.*

This definition of normal modal logics is chosen to make clear the connection with modal algebras. An more common approach is to take define a modal

logics as a set of formulae satisfying conditions **ML1–4**, together with the Rule of Necessitation: if $p \in \mathcal{ML}$ then $\Box p \in \mathcal{ML}$. *Normal* modal logics are then defined as those additionally satisfying the Kripke schema, **K**: $((\Box p \wedge \Box(p \to q)) \to \Box q) \in \mathcal{ML}$. The two specifications are known to be equivalent (see e.g. Chellas, 1980, Ch. 4).

Normal modal logics can be interpreted in terms of the well-known *Kripke semantics*. Specifically, a model $\mathfrak{M}$ of a normal modal logic $\mathcal{ML}$ is a structure $\langle W, R, v\rangle$, where W is a set, R a binary relation on W (i.e. $\langle W, R\rangle$ is a frame), and $v : \mathcal{V} \to 2^W$ a valuation function which is extended over $\mathcal{ML}$ as follows:

$$\begin{aligned} v(\neg\varphi) &= W \setminus v(\varphi) = \{w \in W : w \notin v(\varphi)\}, \\ v(\varphi \wedge \psi) &= v(\varphi) \cap v(\psi), \\ v(\top) &= W, \\ v(\Diamond(\varphi)) &= \{w : (\exists u)[u \in v(\varphi) \text{ and } uRw\}, \end{aligned}$$

for all $\varphi, \psi \in \mathcal{ML}$.

The elements of W are often called *possible worlds*.

6.3.1 Modal logics and algebras. There is an intimate connection between propositional logics and algebras. The set $\mathfrak{F}$ of all modal formulae can be regarded as a *term algebra* (i.e. an absolutely free algebra generated from the propositional constants) by taking the connectives as (syntactic) operators on formulae.

To obtain an algebraic perspective on the structure of a particular modal logic $\mathcal{ML}$ we can construct a *quotient algebra* of $\mathfrak{F}$ relative to the logical equivalence relation of $\mathcal{ML}$. (Given an algebra $\mathfrak{A} = \langle A, f_1, \ldots, f_n\rangle$, and an equivalence relation $\approx$, the quotient algebra of $\mathfrak{A}$ relative to $\approx$ is the structure $\langle A^{\approx}, f_1^{\approx}, \ldots, f_m^{\approx}\rangle$. Let $x^{\approx} = \{y \mid y \approx x\}$. Then $A^{\approx} = \{x^{\approx} \mid x \in A\}$, and $f_i^{\approx}(x_1^{\approx}, \ldots, x_n^{\approx}) = y^{\approx}$ if $f_i(x_1, \ldots, x_n) = y$.) Each element of this quotient algebra will thus correspond to a semantically distinct proposition expressible in the logic.

DEFINITION 3.61 *Given $\mathcal{ML} \subseteq \mathfrak{F}$, the* Lindenbaum-Tarski algebra $\mathfrak{F}_{\mathcal{ML}}$ *of $\mathcal{ML}$ is the quotient algebra of $\mathfrak{F}$ by the equivalence relation*

$$x \approx_{\mathcal{ML}} y \quad \textit{if and only if} \quad x \leftrightarrow y \in \mathcal{ML}. \tag{3.91}$$

The resulting algebras are modal algebras in the sense of Definition 3.7, and thus they are BAOs.

The Lindenbaum-Tarski construction can also be used to characterise the equational class of all modal algebras: each equivalence $x \approx_{\mathcal{ML}} y$ corresponds to a universally quantified equation $\forall v_1, \ldots, v_n[x = y]$, where $v_1, \ldots, v_n$ are all the propositional variables occurring in either x or y.

There is a direct correspondence between modal algebras and modal logics. The rule **ML5** ensures that $\Diamond$ (and hence $\Box$) is functional; **ML6** corresponds

to the algebraic normality condition (3.3) and **ML7** to additivity (3.2). The generality of the correspondence is expressed by the following proposition:

PROPOSITION 3.62 (JÓNSSON, 1993) *Let $\mathcal{V}(\mathcal{ML})$ be the equational class generated by $\mathfrak{F}_{\mathcal{ML}}$. The mapping $\mathcal{ML} \mapsto \mathcal{V}(\mathcal{ML})$ is a dual isomorphism from the lattice of all normal modal logics to the lattice of equational classes of modal algebras.*

6.3.2 $S4$ and interior algebras. One of the better known modal logics is **S4**. This can be defined as a normal modal logic that also satisfies the following axiom schemata:

$$(p \vee \Diamond p) \leftrightarrow \Diamond p, \tag{3.92}$$

$$\Diamond\Diamond p \leftrightarrow \Diamond p. \tag{3.93}$$

S4 is more often defined by the schemata $\Box\varphi \to \varphi$ (**T**) and $\Box\varphi \to \Box\Box\varphi$ (**4**), which are equivalent to those given here.

Clearly, in the Lindenbaum-Tarski algebra generated from the set of theorems of **S4**, these schema will generate equations of the form of 3.4 and 3.5 characterising a closure operator. From a semantic point of view, the class of models of **S4** consists of all Kripke frames whose accessibility relation is reflexive and transitive; thus according to Proposition 3.15, their complex algebras are exactly the closure algebras.

This correspondence is the basis of the encoding of topological relationships into **S4** proposed in Bennett, 1996 and Bennett, 1997 and also investigated in Nutt, 1999. A similar method had previously been used in Bennett, 1994 to encode modal formulae into intuitionistic propositional logic. This is based on the relation of intuitionistic logic to Heyting algebras, which in turn can be interpreted over topological spaces (Sec. 2 above).

Let $\tau \rightleftharpoons \varphi$ denote the one-to-one mapping from terms of closure algebra to syntactically isomorphic formulae of **S4**: specifically φ is obtained from τ by replacing $-$ by $\neg$, $\vee$ by $+$, $\cdot$ by $\wedge$, cl by $\Diamond$ and int by $\Box$. The following Lemma enables us to use deduction in **S4** to determine entailment among equations in closure algebra:

LEMMA 3.63 (BENNETT, 1997) *Let $\tau_1 = 1, \ldots, \tau_n = 1$ be equations of closure algebra and $\tau_i \rightleftharpoons \varphi_i$. Then*

$$\tau_1 = 1, \ldots, \tau_n = 1 \vdash \tau_0 = 1 \quad \textit{iff} \quad \Box\varphi_1, \ldots, \Box\varphi_n \vdash_{S4} \varphi_0. \tag{3.94}$$

Because of Lemma 3.59, we can also use the modal encoding for testing inconsistency of sets of equations and disequations of closure algebra, and hence for reasoning about topological properties and relationships among spatial regions.

6.3.3 A bi-modal spatial logic. We now give a brief overview of a somewhat more expressive modal encoding of topological relationships proposed in Bennett, 1997. This is obtained by employing a bi-modal logic incorporating both **S4** and the 'universal' modal operator (Goranko, V. and Passy, S., 1992), here denoted by $\blacksquare$ (with its dual being denoted, $\blacklozenge$, where $\blacklozenge\varphi \leftrightarrow \neg\blacksquare\neg\varphi$). As before, the $S4$ operators $\Box$ and $\Diamond$ correspond respectively to the as interior and closure operators, int and cl. The interpretation of $\blacksquare\varphi$ is that φ holds at all possible worlds. The universal modal operator is closely related to the better known **S5** modality, which is the logic semantically determined by taking the accessibility relation to be reflexive, symmetric and transitive (i.e. an equivalence relation). However, this allows the possibility that the set of worlds is partitioned into several sets of worlds which are not accessible to each other. Thus, if $\blacksquare$ were an **S5** modality, $\blacksquare\varphi$ it would be true at world w as long as φ holds in all worlds in the same equivalence class as w—not necessarily in all worlds. The axiomatisation of $\blacksquare$ is the same as that of $\Box$, with the addition of the following two schemata (Goranko, V. and Passy, S., 1992; Wolter and Zakharyaschev, 2002):

$$\blacklozenge p \rightarrow \blacksquare\blacklozenge\, p, \tag{3.95}$$

$$\blacksquare p \rightarrow \Box\, p. \tag{3.96}$$

As before, the $S4$ operators $\Box$ and $\Diamond$ correspond respectively to the as interior and closure operators, int and cl. A formula of the form $\blacksquare\varphi$ ensures that the topological condition encoded by φ holds at every point in space. Similarly, $\blacklozenge\varphi$ means that there is some point p satisfying the condition represented by φ. p can be thought of as a *sample point*, which bears *witness* to some topological constraint. For instance, where two regions x and y overlap, the corresponding modal formula $\blacklozenge(\Box(x) \wedge \Box(y))$ ensures the existence of a point which is in the interior of both x and y.

In terms of the bi-modal logic $S5/S4$, a set of key RCC relations are represented as as follows:

$$\mathsf{C}(x,y) \iff \blacklozenge(x \wedge y), \tag{3.97}$$

$$\mathsf{DC}(x,y) \iff \blacksquare(\neg x \vee \neg y), \tag{3.98}$$

$$\mathsf{O}(x,y) \iff \blacklozenge(\Box(x) \wedge \Box(y)), \tag{3.99}$$

$$\mathsf{DR}(x,y) \iff \blacksquare\Diamond(\neg x \vee \neg y), \tag{3.100}$$

$$\mathsf{P}(x,y) \iff \blacksquare(x \rightarrow y), \tag{3.101}$$

$$\neg\mathsf{P}(x,y) \iff \blacklozenge(x \wedge \neg y), \tag{3.102}$$

$$\mathsf{TP}(x,y) \iff \blacksquare(x \rightarrow y) \wedge \blacklozenge(x \wedge c(\neg y)), \tag{3.103}$$

$$\mathsf{NTP}(x,y) \iff \blacksquare(x \rightarrow \Box(y)), \tag{3.104}$$

$$\mathsf{Non\text{-}Empty}(x) \iff \blacklozenge x, \tag{3.105}$$

(3.106) $$\mathsf{Regular}(x) \iff \blacksquare(\Box(\neg x) \vee \Diamond(\Box(x))).$$

All the RCC-8 relations can be expressed in terms of these formula by using conjunction and negation. Further details of how this logic can be used for topological reasoning are given in Bennett, 1997 and Wolter and Zakharyaschev, 2002.

6.4 A proof system for CRAs

In this section we will describe a sound and complete logic for *contact relation algebras*, within which general facts about CRAs can be proved. The semantics of this logic is relational, introduced by Orlowska, 1991 (and further developed in Orlowska, 1996), and the proof system is in the style of Rasiowa and Sikorski, 1963.

Our language $\mathcal{L}$ consists of the disjoint union of the following sets:

i) A set $\{C, 1'\}$ of constants, representing the generating contact relation and the identity.

ii) An infinite set $\mathcal{V}$ of individuum variables.

iii) A set $\{+, \cdot, -, \ ; \ , \breve{}\}$ of names for the relational operators.

iv) A set $\{(,)\}$ of delimiters.

With some abuse of language, we use the same symbols for the actual operations. The terms of the language are defined recursively:

i) C and $1'$ are terms.

ii) If R and S are terms, so are $(R + S)$, $(R \cdot S)$, $(-R)$, $(R; S)$, $(R\breve{})$.

iii) No other string is a term.

The set of all terms will denoted by $\mathcal{T}$. In the sequel, we will follow the usual conventions of reducing brackets. The set $\mathcal{F}$ of $\mathcal{L}$-formulae is

$$\{xRy : R \in \mathcal{T}, x, y \in \mathcal{V}\}.$$

A model of $\mathcal{L}$ is a pair $M = \langle W, m\rangle$, where W is a nonempty set, and $m : \mathcal{T} \to W \times W$ is a mapping such that:

(3.107) $m(C)$ is a contact relation.

(3.108) $m(1')$ is the identity relation on W.

(3.109) m is a homomorphism from the algebra of terms to. $\langle Rel(W), \cup, \cap, -, \ ; \ , \breve{}\rangle$

A valuation v is a mapping from $\mathcal{V}$ to W. If xRy is a formula, then we say that M *satisfies* xRy *under* v, written as $M, v \models xRy$, if $\langle v(x), v(y)\rangle \in m(R)$.

xRy is called *true in the model* M, if $M, v \models xRy$ for all valuations v. xRy is called *valid*, if it is true in all models.

The proof system consists of two types of rules: With *decomposition rules* we can decompose formulae into an equivalent sequence of simpler formulae. The decomposition rules are the same for every system of relation algebras. The *specific rules* are tailored towards the concrete situation; they modify a sequence of formulae and have the status of structural rules. The role of axioms is played by *axiomatic sequences*.

Proofs have the form of trees: Given a formula xRy, we successively apply decomposition or specific rules; in this way we obtain a tree whose root is xRy, and whose nodes consist of sequences of formulae. A branch of a tree is *closed* if it contains a node which contains an axiomatic sequence as a subsequence. A tree is *closed* if all its branches are closed.

$$(\cup) \quad \frac{K, x(R \cup S)y, H}{K, xRy, xSy, H} \qquad (\neg\cup) \quad \frac{K, x - (R \cup S)y, H}{K, x(-R)y, H \mid K, x(-S)y, H}$$

$$(\cap) \quad \frac{K, x(R \cap S), H}{K, xRy, H \mid K, xSy, H} \qquad (\neg\cap) \quad \frac{K, x - (R \cap S)y, H}{K, x(-R)y, x(-S)y, H}$$

$$({}^{\smile}) \quad \frac{K, xR^{\smile}y, H}{K, yRx, H} \qquad (\neg{}^{\smile}) \quad \frac{K, x(-R^{\smile})y, H}{K, y(-R)x, H}$$

$$(\neg-) \quad \frac{K, x(- - R)y, H}{K, xRy, H}$$

$$(\,;\,) \quad \frac{K, x(R\,;\,S)y, H}{K, xRz, H, x(R\,;\,S)y \mid K, zSy, H, x(R\,;\,S)y} \qquad \text{where } z \text{ is any variable}$$

$$(\neg\,;\,) \quad \frac{K, x - (R\,;\,S)y, H}{K, x(-R)z, z(-S)y, H} \qquad \text{where } z \text{ is a restricted variable}$$

Table 6.3. Decomposition rules.

Rasiowa-Sikorski systems are, in way, dual to tableaux: Whereas in the latter one tries to refute the negation of a formula, the Rasiowa-Sikorski systems attempt to verify a formula by closing the branches of a decomposition tree with axiomatic sequences.

The decomposition rules of the system are given in Table 6.3, and the specific rules for the system are given in Table 6.4. There, a variable z is called *restricted* in a rule, if it does not occur in the upper part of that rule. K and H are finite, possibly empty, sequences of $\mathcal{L}$ formulae. For all $R \in \mathcal{T}$, the following

sequences are axiomatic:

(3.110) $$xRy,\ x(-R)y,$$

(3.111) $$x1'x.$$

$$(\mathrm{sym}\,1')\quad \frac{K, x1'y, H}{K, y1'x, H}$$

$$(\mathrm{tran}\,1')\quad \frac{K, x1'y, H}{K, x1'z, H, x1'y \mid K, z1'y, H, x1'y},\quad \text{where } z \text{ is any variable}$$

$$(1'_1)\quad \frac{K, xRy, H}{K, x1'z, H, xRy \mid K, zRy, H, xRy},\quad \text{where } z \text{ is any variable}$$

$$(1'_2)\quad \frac{K, xRy, H}{K, xRz, H, xRy \mid K, z1'y, H, xRy},\quad \text{where } z \text{ is any variable}$$

$$(\mathrm{refl}\,C)\quad \frac{K, xCy, H}{K, x1'y, xCy, H} \qquad (\mathrm{sym}\,C)\quad \frac{K, xCy, H}{K, yCx, H}$$

$$(\mathrm{ext}\,C)\quad \frac{K}{K, x(-C)z, yCz \mid K, y(-C)t, xCt \mid K, x(-1')y}\quad \text{where } z \text{ and } t \text{ are restricted variables}$$

$$(\mathrm{cut}\,C)\quad \frac{K}{K, xCy \mid K, x(-C)y}$$

Table 6.4. Specific rules.

The following result shows that the logic is sound and complete:

PROPOSITION 3.64 (DÜNTSCH AND ORLOWSKA, 2000)

i) *All decomposition rules are admissible.*

ii) *All specific rules are admissible.*

iii) *The axiomatic sequences are valid.*

iv) *If a formula is valid then it has a closed proof tree.*

An example in Düntsch and Orlowska, 2000 shows that there is a CRA with infinitely many atoms below $1'$, and thus, by a result of Andréka et al., 1997, the equational logic of CRAs is undecidable.

7. Conclusion

In this chapter we have examined the topic of region-based spatial representation from a number of perspectives. We have looked at the relationships between algebraic models, point-set topology and axiomatic theories of spatial regions. The approach of modelling space in terms of a Boolean algebra, supplemented with additional operations and/or relations provides a very general and adaptable analysis. Moreover, such algebraic formalisms provide a powerful tool for establishing correspondences between relational axiomatic theories and the models of point-set topology. Specifically, we have seen that Boolean Contact Algebras (BCAs), have essentially the same expressive capabilities as theories such as the Region Connection Calculus (Randell et al., 1992b), which was developed as a knowledge representation formalism for Artificial Intelligence. We have presented representation theorems that characterise topological models of BCAs.

It is interesting to note that the properties of topological spaces that characterise the topological representations of relational theories (according to the representation theorems of Sec. 4) do not coincide with those most familiar to point-set topologists. This is primarily because the elements of region-based theories are modelled as regular subsets of a topological space. Thus, relevant properties for spaces models are typically weaker than better known separation properties, in that they impose conditions on regular subsets of the space rather than on points or on open or closed sets in general. Although we believe that these representations are particularly natural, it is worth noting that there may be alternative topological representations where the embedding of the algebraic structure in the topology takes a different form (as with the representations of Roeper, 1997 and Mormann, 1998).

As well as considering algebras of regions, we have also seen how the formalism of Relation Algebra provides an algebraic treatment of the relational concepts of a theory. This proves to be well suited to representing spatial relations, in that a large vocabulary of significant spatial relations can be equationally defined from just the contact relation, C. The Relation Algebraic analysis also serves to provide a foundation for the technique of compositional inference, which has been found to be effective in a number of AI applications, for reasoning with both temporal (Allen, 1983) and spatial relations (Randell et al., 1992a; Renz and Nebel, 1999).

A technique that has proved particularly useful for reasoning about topology has been the encoding into modal logic. Again, algebra provides a bridging formalism, since propositional logics have a direct correspondence to Boolean algebras with operators. Because the principal function of logical languages is to describe mechanisms of valid inference, much is known about how such inferences can be automated, and about the computational complexity of reasoning

algorithms using these systems. Thus the encodings have led to the development of decision procedures and establishment of complexity results for reasoning about topological relations (Bennett, 1997; Renz and Nebel, 1997; Renz and Nebel, 1999; Wolter and Zakharyaschev, 2000; Wolter and Zakharyaschev, 2002). Modal encodings have also been applied to encode relations in projective geometry (Balbiani et al., 1997; Venema, 1999).

Another promising avenue for extending the use of modal encodings is by the use of multi-dimensional modal logics (Segerberg, 1973; Marx and Venema, 1997; Gabbay et al., 2003), which are multi-modal logics, with different modalities ranging over orthogonal dimensions of their model structures. These have been used to capture both multiple spatial dimensions and the combination of space with time. Yet another approach is to employ modal logics , in which spatial relations are associated with the accessibility relation associated with the modal operators (Cohn, 1993; Lutz and Wolter, 2006).

The current chapter has focused on purely topological aspects of spatial information. However, other geometrical properties have also been treated in terms of region-based relational and algebraic theories. An early paper of Tarski, 1956a showed how the whole of Euclidean geometry could be re-constructed by taking regions (rather than points) as the basic spatial entities, and the relation of parthood and the property of sphericity as the conceptual primitives. A simpler axiomatisation of a theory of this kind is given by Bennett, 2001. Though from a theoretical point of view such formalisms are highly interesting, it is less clear whether they could provide useful mechanisms for computing inferences. It seems that by adding only a little more than topology to a representation one easily obtains a computationally intractable theory. For instance, in Davis et al., 1999 it is shown that reasoning with the RCC-8 relations together with a convexity predicate is already massively intractable. The question of the expressive power of region based theories has been the subject of much research (e.g. Pratt and Schoop, 2000, and Davis, 2006) gives some rather general results demonstrating the very high expressive power of theories that allow quantification over regions.

One potentially very useful development of spatial logics, which has yielded positive results regarding tractability, is the combination of spatial and temporal concepts into a combined spatio-temporal calculus. Certain restricted syntax fragments of modal logics that can encode spatial and temporal information can express a significant range of spatio-temporal relationships whilst remaining tolerably amenable to automated reasoning (Wolter and Zakharyaschev, 2000; Bennett et al., 2002).

An important aspect of space that has not been explicitly considered in this chapter is dimensionality. The formalisms presented in this chapter do not explicitly constrain the dimensionality either of the regions or the embedding space. However, the interpretation of regions as regular sets of a space means

that in such models, regions will all have the same dimension as the whole space. The dimensionality of the space could be fixed by appropriate axioms constraining the connection relation, but the dimensionality of regions would still be uniform. For many applications it would be useful to have a richer theory incorporating regions of different dimensionality into its domain. Axiomatic topological theories that can handle diverse dimensionalities have been proposed in Gotts, 1996, Galton, 1996 and Galton, 1994. However, the relationship between axiomatic theories of this kind and topological models has not been fully established and is certainly a rich area for further work.

The more computationally amenable region-based calculi also suffer from inexpressivity regarding self-connectedness of regions—i.e. the domain can include multi-piece regions, but single and multi-piece regions cannot be distinguished within the theory. (The distinction can easily be made in 1st-order theories such as the full RCC theory, where we can define

$$\forall x \textsf{Self-Connected}(x) \leftrightarrow \forall yz[(x = y + z) \rightarrow yCz,$$

but the full RCC theory is undecidable: Dornheim, 1998.) This is closely related to their inexpressibility in regard to dimensionality. It was shown in Renz, 1998 that any consistent set of RCC-8 relations has a model in which the regions are self-connected regular subsets of a three-dimensional space. However, as explained in Grigni et al., 1995, an interpretation over self-connected regions of two-dimensional space may not be possible, despite the existence of a higher dimensional model. Hence, in this case, enforcing self-connectedness of regions would have no affect on consistency unless we also had some means of enforcing planarity (or linearity) of the space. Developing any kind of computationally effective calculus for reasoning about topological relations between self-connected regions in two-dimensional space has proved elusive. Some results developed from a graph-theoretic viewpoint suggest that this is at least NP-hard, and may well be undecidable (Kratochvíl, 1991; Kratochvíl and Matoušek, 1991).

Another constraint on the structure of space, which has received attention is discreteness. There are many applications, such as describing or reasoning about video images, where one is dealing with a discrete spatial structure. Axiomatic theories which allow atomic regions have been investigated by Masolo and Vieu, 1999, Galton, 1999 and Düntsch and Winter, 2006, while Düntsch and Vakarelov, 2006 present a generalisation of BCAs, in which the extensionality axiom is dropped, and proves a representation theorem in terms of discrete proximity spaces. A deep analysis of the connection between relational, algebraic and topological properties of proximity and contact algebras and their representation spaces, including the discrete case, can be found in Dimov and Vakarelov, 2006.

The diversity of spatial formalisms is testament to the richness and depth of spatial concepts. Indeed Tarski, 1956b suggested that geometrical primitives may provide a conceptual basis from which all precise concepts can be defined. Although axiomatic, region-based theories of topology are increasingly well understood and integrated with related areas of mathematics and knowledge representation, many directions for further research remain open.

Acknowledgments

This work was partially supported by the Engineering and Physical Sciences Research Council of the UK, under grant EP/D002834/1, and by the Natural Sciences and Engineering Research Council of Canada.

References

Ahmed, T. S. (2004). Tarskian algebraic logic. *Journal of Relational Methods in Computer Science*, 1:3–26.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Comm. ACM*, 26(11):832–843.

Andréka, H., Givant, S., and Németi, I. (1997). *Decision problems for equational theories of relation algebras*. Number 604 in Memoirs of the American Mathematical Society. AMS, Providence.

Andréka, H., Németi, I., and Sain, I. (2001). Algebraic logic. In Gabbay, D. M. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 2, pages 133–247. Kluwer, Dordrecht, 2nd edition.

Balbes, R. and Dwinger, P. (1974). *Distributive Lattices*. University of Missouri Press, Columbia.

Balbiani, P., Fariñas del Cerro, L., Tinchev, T., and Vakarelov, D. (1997). Modal logics for incidence geometries. *J. Logic Comput.*, 7(1):59–78.

Bennett, B. (1994). Spatial reasoning with propositional logics. In Doyle et al., 1994.

Bennett, B. (1996). Modal logics for qualitative spatial reasoning. *Bull. IGPL*, 4(1):23–45. Available from `ftp://ftp.mpi-sb.mpg.de/pub/igpl/Journal/V4-1/index.html`.

Bennett, B. (1997). *Logical Representations for Automated Reasoning about Spatial Relationships*. PhD thesis, School of Computing, The University of Leeds. Abstract and postscript at `http://www.scs.leeds.ac.uk/brandon/thesis.html`.

Bennett, B. (2001). A categorical axiomatisation of region-based geometry. *Fund. Inform.*, 46(1–2):145–158.

Bennett, B., Cohn, A. G., Wolter, F., and Zakharyaschev, M. (2002). Multi-dimensional modal logic as a framework for spatio-temporal reasoning. *Applied Intelligence*, 17(3):239–251.

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, Cambridge.

Borgo, S., Guarino, N., and Masolo, C. (1996). A pointless theory of space based on strong connection and congruence. In Aiello, L. C. and Doyle, J., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 5th International Conference (KR96)*. Morgan Kaufmann, San Francisco.

Chellas, B. F. (1980). *Modal Logic: An Introduction*. Cambridge University Press, Cambridge.

Chin, L. and Tarski, A. (1951). Distributive and modular laws in the arithmetic of relation algebras. *University of California Publications in Mathematics*, 1:341–384.

Clarke, B. L. (1981). A calculus of individuals based on 'Connection'. *Notre Dame J. Formal Logic*, 23(3):204–218.

Clarke, B. L. (1985). Individuals and points. *Notre Dame J. Formal Logic*, 26(1):61–75.

Cohn, A. G. (1987). A more expressive formulation of many sorted logic. *J. Automat. Reasoning*, 3:113–200.

Cohn, A. G. (1993). Modal and non-modal qualitative spatial logics. In Anger, F., Guesgen, H. W., and van Benthem, J., editors, *Proceedings of the Workshop on Spatial and Temporal Reasoning at the 13th International Joint Conference on Artificial Intelligence, IJCAI 93*, pages 95–100.

Cohn, A. G., Bennett, B., Gooday, J., and Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 13:1–42.

Cohn, A. G., Gooday, J. M., and Bennett, B. (1994). A comparison of structures in spatial and temporal logics. In Casati, R., Smith, B., and White, G., editors, *Philosophy and the Cognitive Sciences: Proceedings of the 16th International Wittgenstein Symposium*. Hölder-Pichler-Tempsky, Vienna.

Cohn, A. G., Schubert, L., and Shapiro, S., editors (1998). *Principles of Knowledge Representation and Reasoning: Proceedings of the 6th International Conference (KR98))*, San Francisco. Morgan Kaufman.

Davis, E. (2006). The expressivity of quantifying over regions. *J. Logic Comput.* To appear.

Davis, E., Gotts, N., and Cohn, A. G. (1999). Constraint networks of topological relations and convexity. *Constraints*, 4(3):241–280.

de Laguna, T. (1922). Point, line and surface as sets of solids. *J. Philos.*, 19: 449–461.

de Rijke, M. and Venema, Y. (1995). Sahlqvist's theorem for Boolean algebras with operators with an application to cylindric algebras. *Studia Logica*, 54(1):61–78.

Dimov, G. and Vakarelov, D. (2006). Contact algebras and region-based theory of space; a proximity approach, i–ii. *Fundamenta Informaticae*. to appear.

Dornheim, C. (1998). Undecidability of plane polygonal mereotopology. In Cohn et al., 1998, pages 342–353.

Doyle, J. Sandewall, E. and Torasso, P. editors (1994). *Principles of Knowledge Representation and Reasoning: Proceedings of the 4th International Conference (KR94)*, San Francisco. Morgan Kaufmann.

Düntsch, I. (2005). Relation algebras and their application in temporal and spatial reasoning. *Artificial Intelligence Review*, 23:315–357.

Düntsch, I. and Orlowska, E. (2000). A proof system for contact relation algebras. *J. Philos. Logic*, 29:241–262.

Düntsch, I., Schmidt, G., and Winter, M. (2001a). A necessary relation algebra for mereotopology. *Studia Logica*, 69:381–409.

Düntsch, I. and Vakarelov, D. (2006). Region-based theory of discrete spaces: A proximity approach. *Discrete Appl. Math.* To appear.

Düntsch, I., Wang, H., and McCloskey, S. (1999). Relation algebras in qualitative spatial reasoning. *Fund. Inform.*, 39:229–248.

Düntsch, I., Wang, H., and McCloskey, S. (2001b). A relation algebraic approach to the Region Connection Calculus. *Theoret. Comput. Sci. (B)*, 255:63–83.

Düntsch, I. and Winter, M. (2004). Construction of Boolean contact algebras. *AI Commun.*, 13:235–246.

Düntsch, I. and Winter, M. (2005). A representation theorem for Boolean contact algebras. *Theoret. Comput. Sci. (B)*, 347:498–512.

Düntsch, I. and Winter, M. (2006). Remarks on lattices of contact relations. Research report, Department of Computer Science, Brock University.

Efremovič, V. A. (1952). The geometry of proximity I. *Mat Sbornik (New Series)*, 31:189–200. In Russian.

Egenhofer, M. (1991). Reasoning about binary topological relations. In Gunther, O. and Schek, H. J., editors, *Proceedings of the Second Symposium on Large Spatial Databases, SSD'91 (Zurich, Switzerland)*, volume 525 of *Lect. Notes Comp. Sci.*, pages 143–160. Springer Verlag, Heidelberg.

Egenhofer, M. and Franzosa, R. (1991). Point–set topological spatial relations. *International Journal of Geographic Information Systems*, 5(2):161–174.

Egenhofer, M. and Herring, J. (1991). Categorizing binary topological relationships between regions, lines and points in geographic databases. Tech. report, Department of Surveying Engineering, University of Maine.

Engelking, Ryszard (1977). *General Topology*. PWN—Polish Scientific Publishers, Warszawa.

Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54:199–227.

Freksa, C. and Mark, D. M., editors (1999). *Spatial information theory: Cognitive and computational foundations of geographic information science — Proceedings of COSIT'99*, number 1661 in Lect. Notes Comp. Sci., Heidelberg. Springer Verlag.

Gabbay, D., Kurucz, A., Wolter, F., and Zakharyaschev, M. (2003). *Many-Dimensional Modal Logics: Theory and Applications*, volume 148 of *Studies in Logic and the Foundations of Mathematics*. North–Holland, Amsterdam.

Galton, A. P. (1994). Multidimensional mereotopology. In Doyle et al., 1994.

Galton, A. P. (1996). Taking dimension seriously in qualitative spatial reasoning. In Wahlster, W., editor, *Proceedings of the 12th European Conference on Artificial Intelligence*, pages 501–505. Wiley, Chichester.

Galton, A. P. (1999). The mereotopology of discrete space. In Freksa and Mark, 1999, pages 251–266.

Givant, S. (2006). The calculus of relations as a foundation for mathematics. *J. Automat. Reasoning*. To appear.

Goodman, N. (1951). *The Structure of Appearance*. Bobbs-Merill (second edition, 1966), Indianapolis.

Goranko, V. and Passy, S. (1992). Using the universal modality: Gains and questions. *J. Logic Comput.*, 2:5–30.

Gotts, N. M. (1996). Formalising commonsense topology: The INCH calculus. In Kautz, H. and Selman, B., editors, *Proc. of the Fourth International Symposium on Artificial Intelligence and Mathematics*, pages 72–75.

Grigni, M., Papadias, D., and Papadimitriou, C. (1995). Topological inference. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, pages 901–907. Morgan Kaufmann, San Francisco.

Hayes, P. J. (1985). The second naive physics manifesto. In Hobbs, J. R. and Moore, B. editors, *Formal Theories of the Commonsense World*, pages 1–36. Ablex Publishing Corp., Norwood.

Heinrich, W. (1978). *Discriminator–Algebras*, volume 6 of *Studien zur Algebra und ihre Anwendungen*. Akademie Verlag, Berlin.

Henkin, L., Monk, J. D., and Tarski, A. (1971). *Cylindric algebras, Part I*. North–Holland, Amsterdam.

Henkin, L., Monk, J. D., and Tarski, A. (1985). *Cylindric algebras, Part II*. North–Holland, Amsterdam.

Hirsch, R. and Hodkinson, I. (2002). *Relation algebras by games*, volume 147 of *Studies in Logic and the Foundations of Mathematics*. North–Holland, Amsterdam.

Hughes, G. E. and Cresswell, M. J. (1968). *An Introduction to Modal Logic*. Methuen, London.

Johnstone, P. T. (1983). The point of pointless topology. *Bull. Amer. Math. Soc.*, 8:41–53.

Jónsson, B. (1991). The theory of binary relations. In Andréka, H., Monk, J. D., and Németi, I., editors, *Algebraic Logic*, volume 54 of *Colloquia Mathematica Societatis János Bolyai*, pages 245–292. North Holland, Amsterdam.

Jónsson, B. (1993). A survey of Boolean algebras with operators. In Rosenberg, I. G. and Sabidussi, G., editors, *Algebras and Orders*, volume 389 of *NATO Adv. Sci. Inst. Ser. C, Math. Phys. Sci.*, pages 239–286. Kluwer, Dordrecht.

Jónsson, B. (1994). On the canonicity of Sahlqvist identities. *Studia Logica*, 53:473–491.

Jónsson, B. (1995). The preservation theorem for canonical extensions of Boolean algebras with operators. In Baker, K.A. and Wille, R., editors, *Lattice theory and its applications*, pages 121–130. Heldermann, Lemgo.

Jónsson, B., Andréka, H., and Németi, I. (1991). Free algebras in discriminator varieties. *Algebra Universalis*, 28:401–447.

Jónsson, B. and Tarski, A. (1951). Boolean algebras with operators I. *Amer. J. Math.*, 73:891–939.

Koppelberg, S. (1989). *General Theory of Boolean Algebras*, volume 1 of *Handbook of Boolean Algebras*. North–Holland, Amsterdam.

Kratochvíl, J. (1991). String graphs II: Recognizing string graphs is NP-hard. *J. Combin. Theory Ser. B*, 52:67–78.

Kratochvíl, J. and Matoušek, J. (1991). String graphs requiring exponential representations. *J. Combin. Theory Ser. B*, 53:1–4.

Leonard, H. S. and Goodman, N. (1940). The calculus of individuals and its uses. *J. Symbolic Logic*, 5:45–55.

Leśniewski, S. (1927–1931). O podstawach matematyki. *Przeglad Filozoficzny*, 30–34.

Leśniewski, S. (1983). On the foundation of mathematics. *Topoi*, 2:7–52.

Leśniewski, S. (1992). Foundations of the general theory of sets I (1916). In Surma, S. J., Srzednicki, J., Barnett, D. I., and Rickey, F. V., editors, *S. Leśniewski: Collected Works*, volume 1, pages 129–173. Kluwer, Dordrecht.

Li, S. and Ying, M. (2003a). Extensionality of the RCC8 composition table. *Fund. Inform.*, 55(3–4):363–385.

Li, S. and Ying, M. (2003b). Region Connection Calculus: Its models and composition table. *Artificial Intelligence*, 145(1–2):121–146.

Li, S., Ying, M., and Li, Y. (2005). On countable RCC models. *Fund. Inform.*, 62:329–351.

Lutz, C. and Wolter, F. (2006). Modal logics of topological relations. *Logical Methods in Computer Science*. To appear.

Madarász, J. X. (1998). Interpolation in algebraizable logics; semantics for non-normal multi-modal logic. *J. Appl. Non–Classical Logics*, 8:67–105.

Marx, M. and Venema, Y. (1997). *Multi-Dimensional Modal Logic*. Applied Logic. Kluwer, Dordrecht.

Masolo, C. and Vieu, L. (1999). Atomicity vs. infinite divisibility of space. In Freksa and Mark, 1999, pages 235–250.

McKinsey, J. C. C. and Tarski, A. (1944). The algebra of topology. *Ann. of Math.*, 45:141–191.

Mormann, T. (1998). Continuous lattices and Whiteheadian theory of space. *Logic Log. Philos.*, 6:35–54.

Naimpally, S. A. and Warrack, B. D. (1970). *Proximity Spaces*. Cambridge University Press, Cambridge.

Németi, I. (1991). Algebraizations of quantifier logics: An introductory overview. *Studia Logica*, 50(3–4):485–569. Special issue on Algebraic Logic (eds. W. J. Block and D. Pigozzi).

Nicod, J. (1924). *Geometry in the Sensible World*. Doctoral thesis, Sorbonne. English translation in *Geometry and Induction*, Routledge and Kegan, London, 1969.

Nutt, W. (1999). On the translation of qualitative spatial reasoning problems into modal logics. In Burgard, W., Christaller, T., and Cremers, A. B., editors, *Proceedings of the 23rd Annual German Conference on Advances in Artificial Intelligence (KI-99)*, volume 1701 of *Lect. Notes AI*, pages 113–124. Springer Verlag, Heidelberg.

Orlowska, E. (1991). Relational interpretation of modal logics. In Andréka, H., Monk, J. D., and Németi, I., editors, *Algebraic Logic*, volume 54 of *Colloquia Mathematica Societatis János Bolyai*, pages 443–471. North Holland, Amsterdam.

Orlowska, E. (1996). Relational proof systems for modal logics. In Wansing, H., editor, *Proof theory of modal logic*, pages 55–78. Kluwer, Dordrecht.

Pratt, I. and Lemon, O. (1997). Ontologies for plane, polygonal mereotopology. *Notre Dame J. Formal Logic*, 38(2):225–245.

Pratt, I. and Schoop, D. (1998). A complete axiom system for polygonal mereotopology of the real plane. *J. Philos. Logic*, 27(6):621–661.

Pratt, I. and Schoop, D. (2000). Expressivity in polygonal, plane mereotopology. *J. Symbolic Logic*, 65:822–838.

Pratt-Hartmann, I. (2001). Empiricism and rationalism in region-based theories of space. *Fund. Inform.*, 46:159–186.

Randell, D. A., Cohn, A. G., and Cui, Z. (1992a). Computing transitivity tables: A challenge for automated theorem provers. In Kapur, D., editor, *Proceedings of the 11th International Conference on Automated Deduction (CADE-11)*, volume 607 of *Lect. Notes AI*, pages 786–790. Springer Verlag, Heidelberg.

Randell, D. A., Cui, Z., and Cohn, A. G. (1992b). A spatial logic based on regions and connection. In *Principles of Knowledge Representation and Reasoning:*

Proceedings of the 3rd International Conference (KR92), pages 165–176. Morgan Kaufmann, San Francisco.

Rasiowa, H. and Sikorski, R. (1963). *The Mathematics of Metamathematics*, volume 41 of *Polska Akademia Nauk. Monografie matematyczne*. PWN—Polish Scientific Publishers, Warsaw.

Renz, J. (1998). A canonical model of the Region Connection Calculus. In Cohn et al., 1998, pages 330–341.

Renz, J. and Nebel, B. (1997). On the complexity of qualitative spatial reasoning: a maximal tractable fragment of the Region Connection Calculus. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97*, pages 522–527.

Renz, J. and Nebel, B. (1999). On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the Region Connection Calculus. *Artificial Intelligence*, 108(1–2):69–123.

Roeper, P. (1997). Region based topology. *J. Philos. Logic*, 26:251–309.

Röhrig, R. (1994). A theory for qualitative spatial reasoning based on order relations. In *Proceedings of the 12th National Conference on Artificial Intelligence*, volume 2, pages 1418–1424. MIT Press, Menlo Park.

Segerberg, K. (1973). Two-dimensional modal logic. *J. Philos. Logic*, 2:77–96.

Shchepin, E. V. (1972). Real-valued functions and spaces close to normal. *Siberian Math. J.*, 13:820–830.

Stell, J. G. (2000). Boolean connection algebras: A new approach to the Region Connection Calculus. *Artificial Intelligence*, 122:111–136.

Tarski, A. (1929). Les fondaments de la géométrie des corps. Annales de la Société Polonaise de Mathématique, pages 29–33.

Tarski, A. (1941). On the calculus of relations. *J. Symbolic Logic*, 6:73–89.

Tarski, A. (1956a). Foundations of the geometry of solids (1929). In Woodger, 1956, pages 29–33.

Tarski, A. (1956b). Some methodological investigations on the definability of concepts. In Woodger, 1956, pages 296–319.

Tarski, A. (1959). What is elementary geometry? In Henkin, L., Suppes, P., and Tarski, A., editors, *The Axiomatic Method (with special reference to geometry and physics)*, pages 16–29. North–Holland, Amsterdam.

Tarski, A. and Givant, S. (1987). *A formalization of set theory without variables*, volume 41 of *Colloquium Publications*. AMS, Providence.

Tarski, A. and Givant, S. (1999). Tarski's system of geometry. *Bull. Symbolic Logic*, 5(2):175–214.

Vakarelov, D., Dimov, G., Düntsch, I., and Bennett, B. (2002). A proximity approach to some region–based theories of space. *J. Appl. Non–Classical Logics*, 12:527–529.

Vakarelov, D., Düntsch, I., and Bennett, B. (2001). A note on proximity spaces and connection based mereology. In Welty, C. and Smith, B., editors, *Pro-*

ceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS'01), pages 139–150. ACM.

van Benthem, J. (1984). Correspondence theory. In Gabbay, D. M. and Guenthner, F., editors, *Extensions of classical logic*, volume 2 of *Handbook of Philosophical Logic*, pages 167–247. Reidel, Dordrecht.

Venema, Y. (1999). Points, lines and diamonds: A two-sorted modal logic for projective planes. *J. Log. Comput.*, 9(5):601–621.

Vilain, M. and Kautz, H. (1986). Constraint propagation algorithms for temporal reasoning. In Kehler, T. and Rosenschein, S., editors, *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 377–382. MIT Press, Menlo Park.

Whitehead, A. N. (1920). *The Concept of Nature*. Cambridge University Press, Cambridge.

Whitehead, A. N. (1929). *Process and Reality*. MacMillan, New York. A corrected edition of this was published in 1978 by Macmillan.

Whitehead, A. N. (1978). *Process and Reality (corrected edition)*. Macmillan, New York. This is a revised version of a 1929 edition edited by D.R. Griffin and D.W. Sherburne.

Wolter, F. and Zakharyaschev, M. (2000). Spatio-temporal representation and reasoning based on RCC-8. In *Principles of Knowledge Representation and Reasoning: Proceedings of the 7th International Conference (KR00)*, pages 3–14. Morgan Kaufman, San Francisco.

Wolter, F. and Zakharyaschev, M. (2002). Qualitative spatio-temporal representation and reasoning: A computational perspective. In Lakemeyer, G. and Nebel, B., editors, *Exploring Artificial Intelligence in the New Millenium*, pages 175–216. Morgan Kaufmann, San Francisco.

Woodger, J. H., editor (1956). *Logic, Semantics, Metamathematics*, Oxford. Clarendon Press.

Chapter 4

QUALITATIVE SPATIAL REASONING USING CONSTRAINT CALCULI

Jochen Renz
Australian National University

Bernhard Nebel
Albert-Ludwig University of Freiburg

Second Reader

Ian Pratt-Hartmann
University of Manchester

1. Introduction

Qualitative reasoning is an approach for dealing with commonsense knowledge without using numerical computation. Instead, one tries to represent knowledge using a limited vocabulary such as qualitative relationships between entities or qualitative categories of numerical values, for instance, using $\{+, -, 0\}$ for representing real values. An important motivation for using a qualitative approach is that it is considered to be closer to how humans represent and reason about commonsense knowledge. Another motivation is that it is possible to deal with incomplete knowledge. Qualitative reasoning, however, is different from fuzzy computation. While fuzzy categories are approximations to real values, qualitative categories make only as many distinctions as necessary—the granularity depends on the corresponding application.

Two very important concepts of commonsense knowledge are time and space. Time, being a scalar entity, is very well suited for a qualitative approach and, thus, qualitative temporal reasoning has early emerged as a lively sub-field of qualitative reasoning which has generated a lot of research effort and important results. Space, in turn, is much more complex than time. This is mainly due

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 161–215.
© 2007 *Springer.*

to its inherent multi-dimensionality which leads to a higher degree of freedom and an increased possibility of describing entities and relationships between entities. This becomes clear when enumerating natural language expressions involving space or time. While temporal expressions mainly describe order and duration (like "before", "during", "long", or "a while") or a personal or general temporal category (like "late" or "morning"), spatial expressions are manifold. They are used for describing, for instance, direction ("left", "above"), distance ("far","near"), size ("large", "tiny"), shape ("oval", "convex"), or topology ("touch", "inside"). It is obvious that most spatial expressions in natural language are purely qualitative.

Although there are doubts that because of its multi-dimensionality, space can be adequately dealt with by using only qualitative methods (the poverty conjecture of Forbus et al., 1987), qualitative spatial reasoning has become an active research area. Because of the richness of space and its multiple aspects, however, most work in qualitative spatial reasoning has focused on single aspects of space. The most important aspects of space are topology, orientation, and distance. As shown in psychological studies (Piaget and Inhelder, 1948), this is also the order in which children acquire spatial notions. Other aspects of space include size, shape, morphology, and spatial change (motion).

Orthogonal to this view is the question for the right spatial ontology. One line of research considers points as the basic entities, another line considers extended spatial entities such as spatial regions as basic entities. While it is easier to deal with points rather than with regions in a computational framework, taking regions as the basic entities is certainly more adequate for commonsense reasoning—eventually, all physical objects are extended spatial entities. Furthermore, if points are required, they can be constructed from regions (Biacino and Gerla, 1991). A further ontological distinction is the nature of the embedding space. The most common notion of space is n-dimensional continuous space ($\mathbb{R}^n$). But there are also approaches which consider, e.g., discrete (Galton, 1999) or finite space (Gotts, 1996).

The most popular reasoning methods used in qualitative spatial reasoning are constraint based techniques adopted from previous work in temporal reasoning (see Sec. 2 for a comprehensive introduction to these techniques). In this chapter, we will focus exclusively on these techniques. In order to apply them, it is necessary to have a set of qualitative binary *basic relations* which have the property of being jointly exhaustive and pairwise disjoint, i.e., between any two spatial entities exactly one of the basic relations holds. The set of all possible relations is then the set of all possible unions of the basic relations. Reasoning can be done by exploiting *composition* of relations. For instance, if the binary relation R_1 holds between entities A and B and the binary relation R_2 holds between B and C, then the composition of R_1 and R_2 restricts the

possible relationship between A and C. Compositions of relations are usually pre-computed and stored in a *composition table*.

The rest of the chapter is structured as follows. In Sec. 2, we explain the general idea of constraint-based reasoning. In Sec. 3, a number of different constraint calculi are introduced, which cover topology, orientation, and distance. Since we are interested in reasoning in these spatial calculi, we have to consider what computational resources are necessary to accomplish that. For this purpose, we will introduce computational complexity theory and explain how it can be applied in the context of constraint based reasoning. As we will see, almost all qualitative spatial calculi are computationally intractable. However, it is possible to identify tractable subsets, as we will show in Sec. 5. Based in these results, we will have a look at the *practical efficiency* of reasoning in spatial calculi using tractable fragments in Sec. 6. Finally, in Sec. 7, we consider the combination of different spatial calculi.

2. Constraint-based methods for qualitative spatial representation and reasoning

Knowledge about entities or about the relationships between entities is often given in the form of *constraints*. For instance, when trying to place furniture in a room there are certain constraints on the position of the objects. Unary constraints such as "The room is 5 metres in length and 6 in breadth" restrict the domain of single variables, the length and the breadth of the room. Binary constraints like "The desk should be placed in front of the window", ternary constraints like "The table should be placed between the sofa and the armchair", or in general n-ary constraints restrict the domain of 2, 3, or n variables. Problems like these can be formalised as constraint satisfaction problems.

Given a set of m variables $\mathcal{V} = \{x_1, \ldots, x_m\}$ over a domain $\mathcal{D}$, an *n-ary constraint* consists of an n-ary relation $R_i \subseteq \mathcal{D}^n$ and an n-tuple of variables $\langle x_{i_1}, \ldots, x_{i_n} \rangle$, written $R_i(x_{i_1}, \ldots, x_{i_n})$. For binary constraints, we will also use the infix notation $x_1 R_i x_2$. A (partial) *instantiation* f of variables to values is a (partial) function from the set of variables $\mathcal{V}$ to the set of values $\mathcal{D}$. We say that an instantiation f *satisfies the constraint* $R_i(x_{i_1}, \ldots, x_{i_n})$ if and only if $\langle f(x_{i_1}), \ldots, f(x_{i_n}) \rangle \in R_i$.

A *constraint satisfaction problem* (CSP) consists of a set of variables $\mathcal{V}$ over a domain $\mathcal{D}$ and a set of constraints Θ. The intention is to find a *solution* which is an instantiation such that all constraints in Θ are satisfied.

In this work we restrict ourselves to binary CSPs, i.e., CSPs where only binary constraints are used. A binary CSP can be represented by a *constraint network* which is a labelled digraph where each node is labelled by a variable x_i or by the variable index i and each directed edge is labelled by a binary relation. We will use the notation R_{ij} to denote the relation constraining the

variable pair $\langle x_i, x_j \rangle$. By overloading notation, we also use R_{ij} to denote the constraint $R_{ij}(x_i, x_j)$ itself.

A CSP is *consistent* if it has a *solution*. If the domain of the variables is finite, CSPs can be solved by *backtracking* over the ordered domains of the single variables. Backtracking works by successively instantiating variables with values of the ordered domain until either all variables are instantiated and a solution is found or an inconsistency is detected in which case the current variable is instantiated with the next value of its domain. If all possible instantiations of the current variable lead to an inconsistency, the previous variable becomes the current variable and the process is repeated. Backtracking is in general exponential in the number of variables. The process can be sped up by propagating constraints between the variables and eliminating impossible values as soon as possible. If the domain of the variables is infinite, backtracking over the domain is not possible and other methods have to be applied.

2.1 Binary Constraint Satisfaction Problems and Relation Algebras

One way of dealing with infinite domains is using constraints over a finite set of binary relations. Ladkin and Maddux (1994) proposed to employ *relation algebras* developed by Tarski, 1941 for this purpose. A relation algebra consists of a set of binary relations $\mathcal{R}$ which is closed under several operations on relations and contains some particular relations. The operations are *union* ($\cup$), *intersection* ($\cap$), *composition* ($\circ$), *complement* ($\bar{\cdot}$), and *conversion* ($\cdot^{\smile}$), where conversion and composition are defined as follows:

$$R \circ S \stackrel{def}{=} \{\langle x, y \rangle \mid \exists z \colon \langle x, z \rangle \in R \wedge \langle z, y \rangle \in S\} \tag{4.1}$$

$$R^{\smile} \stackrel{def}{=} \{\langle x, y \rangle \mid \langle y, x \rangle \in R\} \tag{4.2}$$

In the following, we will—by abusing notation—identify sets of relations with their union. For example, we identify $\{R, S, T\}$ with $R \cup S \cup T$.

The particular binary relations mentioned above are the *empty relation* $\emptyset$ which does not contain any pair, the *universal relation* $*$ which contains all possible pairs, and the *identity relation* Id which contains all pairs of identical elements.

We assume that a set of constraints Θ contains one constraint for each pair of variables involved in Θ, i.e., if no information is given about the relation holding between two variables x_i and x_j, then the universal relation $*$ constrains the pair, i.e., $R_{ij} = *$. Another assumption that we make is that whenever a constraint R_{ij} between x_i and x_j is in Θ, the converse relation constrains x_j and x_i, i.e., $(R_{ij})^{\smile} = R_{ji}$.

Determining consistency for CSPs with infinite domains is in general undecidable (Hirsch, 1999). A partial method for determining inconsistency of a CSP is the *path-consistency method* which enforces path-consistency of a CSP (Montanari, 1974; Mackworth, 1977). A CSP is *path-consistent* if and only if for any partial instantiation of any two variables satisfying the constraints between the two variables, it is possible for any third variable to extend the partial instantiation to this third variable satisfying the constraints between the three variables.

A straight-forward way to enforce path-consistency on a binary CSP is to strengthen relations by successively applying the following operation until a fixed point is reached:

$$\forall k : R_{ij} := R_{ij} \cap (R_{ik} \circ R_{kj}).$$

The resulting CSP is *equivalent* to the original CSP in the sense that it has the same set of solutions. If the empty relation results while performing this operation, we know that the CSP is inconsistent. Otherwise, the CSP might or might not be consistent. Provided that the composition of relations can be computed in constant time, the algorithm sketched has a running time of $O(n^5)$, where n is the total number of nodes in the graph. More advanced algorithms can enforce path-consistency in time $O(n^3)$ (Mackworth and Freuder, 1985). Fig. 4.1 shows the $O(n^3)$ time path-consistency algorithm by van Beek, 1992, which uses a queue to keep track of those triples of variables that might be affected by the changes made and which have to be analysed again.

2.2 Relation Algebras based on JEPD Relations

Of particular interest are relation algebras that are based on finite sets of *jointly exhaustive and pairwise disjoint* (JEPD) relations. JEPD relations are sometimes called *atomic*, *basic*, or *base relations*. We refer to them as basic relations. Since any two entities are related by exactly one of the basic relations, they can be used to represent definite knowledge with respect to the given level of granularity. Indefinite knowledge can be specified by unions of possible basic relations. In this chapter we denote a set of basic relations with $\mathcal{B}$ and it should be clear from the context which particular set of basic relations we are referring to. If the set of relations formed by generating all unions over these basic relations is closed under composition and converse, then this set of relations is the carrier of a relation algebra. We denote the set of all relations by $2^{\mathcal{B}}$ alluding to the fact that we identify sets of relations with their unions.

For these relation algebras, the universal relation is the union over all basic relations. Converse, complement, intersection and union of relations can easily be obtained by performing the corresponding set theoretic operations. Composition of basic relations has to be computed using the semantics of the relations.

Algorithm: PATH-CONSISTENCY
Input: A set Θ of binary constraints over the variables $x_1, x_2, \ldots, x_n$
Output: path-consistent set equivalent to Θ, or *fail*, if inconsistency is detected

1. Q $\leftarrow \{(i,j,k),(k,i,j) \mid i < j, k \neq i, k \neq j\}$;
   (i indicates the i-th variable of Θ. Analogously for j and k)
2. *while* Q $\neq \emptyset$ *do*
3. select and delete a path (i,k,j) from Q;
4. *if* REVISE(i,k,j) *then*
5. *if* $R_{ij} = \emptyset$ *then* return `fail`
6. *else* Q $\leftarrow$ Q $\cup\{(i,j,k),(k,i,j) \mid k \neq i, k \neq j\}$;

Function: REVISE(i,k,j)
Input: three variables i, k and j
Output: true, if R_{ij} is revised; false otherwise.
Side effects: R_{ij} and R_{ji} revised using the operations $\cap$ and $\circ$
over the constraints involving i, k, and j.

1. oldR := R_{ij};
2. R_{ij} := $R_{ij} \cap (R_{ik} \circ R_{kj})$;
3. *if* (oldR = R_{ij}) *then* return *false*;
4. R_{ji} := $\breve{R_{ij}}$;
5. return *true*.

Figure 4.1. Van Beek's PATH-CONSISTENCY algorithm.

Composition of unions of basic relations can be obtained by computing the union of the composition of the basic relations. Usually, compositions of the basic relations are pre-computed and stored in a *composition table*.

The best known example of such a relation algebra is the *Interval Algebra* introduced by Allen, 1983, which defines 13 different basic relations between convex intervals on a directed line. The basic relations and a graphical depiction are given in Table 4.1. Even though the interval algebra was introduced for temporal representation and reasoning, there is a number of spatial calculi which are derived from the interval algebra. Some of them we will mention in this chapter.

We say that a relation R is a *refinement* of a relation S if and only if $R \subseteq S$. Given, for instance, a union of relations $\{R_1, R_2, R_3\}$, then the relation $\{R_1, R_2\}$ is a refinement of the former relation. This definition carries over to constraints and to sets of constraints. Then, a set of constraints Θ' is a refinement of Θ if and only if both CSPs have the same variables and for all relations R'_{ij}

Interval Base Relation	Symbol	Pictorial Example	Endpoint Relations
x before y	$\prec$	xxx	$X^- < Y^-, \quad X^- < Y^+,$
y after x	$\succ$	yyy	$X^+ < Y^-, \quad X^+ < Y^+$
x meets y	m	xxxx	$X^- < Y^-, \quad X^- < Y^+,$
y met-by x	$\mathsf{m}^{\smile}$	yyyy	$X^+ = Y^-, \quad X^+ < Y^+$
x overlaps y	o	xxxx	$X^- < Y^- \quad X^- < Y^+,$
y overlapped-by x	$\mathsf{o}^{\smile}$	yyyy	$X^+ > Y^-, \quad X^+ < Y^+$
x during y	d	xxx	$X^- > Y^-, \quad X^- < Y^+,$
y includes x	$\mathsf{d}^{\smile}$	yyyyyyyy	$X^+ > Y^-, \quad X^+ < Y^+$
x starts y	s	xxx	$X^- = Y^-, \quad X^- < Y^+,$
y started-by x	$\mathsf{s}^{\smile}$	yyyyyyyy	$X^+ > Y^-, \quad X^+ < Y^+$
x finishes y	f	xxx	$X^- > Y^-, \quad X^- < Y^+,$
y finished-by x	$\mathsf{f}^{\smile}$	yyyyyyyy	$X^+ > Y^-, \quad X^+ = Y^+$
x equals y	$\equiv$	xxxx yyyy	$X^- = Y^-, \quad X^- < Y^+,$ $X^+ > Y^-, \quad X^+ = Y^+$

Table 4.1. The thirteen basic relations of Allen's interval algebra.

constraining the pair x_i, x_j in Θ' and all relations R_{ij} constraining the same variables in Θ, we have $R'_{ij} \subseteq R_{ij}$. Θ' is said to be a *consistent refinement* of Θ if and only if Θ' is a refinement of Θ and both Θ and Θ' are consistent. A *consistent scenario* Θ_s of a set of constraints Θ is a consistent refinement of Θ where all the constraints of Θ_s are assertions of basic relations.

This chapter deals with determining consistency of binary constraint satisfaction problems that are based on JEPD relations. Let $\mathcal{B}$ be a finite set of JEPD binary relations. The *consistency problem* $\mathsf{CSPSAT}(\mathcal{S})$ for sets $\mathcal{S} \subseteq 2^{\mathcal{B}}$ over a (possibly infinite) domain $\mathcal{D}$ is defined as follows:

Instance: A set $\mathcal{V}$ of variables over a domain $\mathcal{D}$ and a finite set Θ of binary constraints $R(x_i, x_j)$, where $R \in \mathcal{S}$ and $x_i, x_j \in \mathcal{V}$.

Question: Is there an instantiation of all variables in Θ with values from $\mathcal{D}$ such that all constraints are satisfied?

In the general case CSPSAT is undecidable (Hirsch, 1999), but in many interesting cases it is possible to prove decidability or even tractability of $\mathsf{CSPSAT}(\mathcal{S})$.

If CSPSAT is decidable for a certain subset $\mathcal{S} \subseteq 2^{\mathcal{B}}$ then it is possible to decide CSPSAT for other subsets of $2^{\mathcal{B}}$ by using a non-deterministic algorithm. This is done by selecting refinements of the relations such that the refinements are contained in $\mathcal{S}$. For example, suppose $\mathcal{S} \subseteq 2^{\mathcal{B}}$ contains the relations $S_1, \ldots, S_n$, the relation $R \in 2^{\mathcal{B}}$ is not contained in $\mathcal{S}$, and $R = S_1 \cup S_3 \cup S_4$. Then, the constraint $R(x, y)$ can be processed by guessing non-deterministically one of the relations S_1, S_3, and S_4. In general, a subset $\mathcal{S} \subseteq 2^{\mathcal{B}}$ *splits* another subset $\mathcal{T} \subseteq 2^{\mathcal{B}}$ *exhaustively* if for every relation T of $\mathcal{T}$ there are refinements

Algorithm: CONSISTENCY
Input: A set Θ of binary constraints over the variables $x_1, x_2, \ldots, x_n$ and a subset $\mathcal{S} \subseteq 2^{\mathcal{B}}$ that splits $2^{\mathcal{B}}$ exhaustively and for which $\mathsf{CSPSAT}(\mathcal{S})$ is decidable.
Output: `true`, iff Θ is consistent.

1. PATH-CONSISTENCY(Θ)
2. *if* Θ contains the empty relation *then* return `false`
3. *else* choose an unprocessed constraint $R(x, y)$ and
   split R into $S_1, \ldots, S_k \in \mathcal{S}$ such that $S_1 \cup \ldots \cup S_k = R$
4. *if* no constraint can be split *then* return DECIDE(Θ)
5. *for* all refinements S_l $(1 \leq l \leq k)$ *do*
6. replace $R(x, y)$ with $S_l(x, y)$ in Θ
7. *if* CONSISTENCY(Θ) *then* return `true`

Figure 4.2. Backtracking algorithm for deciding consistency.

$S_1, \ldots, S_k \in \mathcal{S}$ such that $T = S_1 \cup \ldots \cup S_k$. If $\mathcal{S}$ splits $\mathcal{T}$ exhaustively, it is obvious that decidability of $\mathsf{CSPSAT}(\mathcal{S})$ implies decidability of $\mathsf{CSPSAT}(\mathcal{T})$. Furthermore, it implies that $\mathsf{CSPSAT}(\mathcal{T})$ can be decided in polynomial time on a non-deterministic Turing machine, if $\mathsf{CSPSAT}(\mathcal{S})$ can be decided in polynomial time.

The non-deterministic algorithm sketched above can be turned into a deterministic one by employing a backtracking scheme. The backtracking algorithm given in Fig. 4.2 is a generalisation of the one proposed by Ladkin and Reinefeld, 1992, and relies on a set $\mathcal{S}$ that splits $2^{\mathcal{B}}$ exhaustively. Note that a set $\mathcal{S}$ splits $2^{\mathcal{B}}$ exhaustively if and only if $\mathcal{S}$ contains all base relations. $\mathcal{S}$ is called the *split set*. The backtracking algorithm uses a function DECIDE which is a sound and complete decision procedure for $\mathsf{CSPSAT}(\mathcal{S})$. The (optional) procedure PATH-CONSISTENCY in line 1 is used as forward-checking and restricts the remaining search space. Nebel, 1997 showed that this restriction preserves soundness and completeness of the algorithm—provided the split set is closed under intersection, composition, and converse. If the decision procedure DECIDE runs in polynomial time, CONSISTENCY is exponential in the number of constraints of Θ. If enforcing path-consistency is sufficient for deciding $\mathsf{CSPSAT}(\mathcal{S})$, DECIDE($\Theta$) in line 4 is not necessary and one can return `true` at this point.

The efficiency of the backtracking algorithm depends on several factors. One of them is, of course, the size of the search space which has to be explored. A common way of measuring the size of the search space is the average *branching factor*, i.e., the average number of branches each node in the search space has. For the backtracking algorithm described in Fig. 4.2 this depends on the average

number of relations of the split set $\mathcal{S}$ into which a relation has to be split. The less splits in average the better, i.e., it is to be expected that the efficiency of the backtracking algorithm depends on the split set $\mathcal{S}$ and its branching factor. Another factor is how the search space is explored. The backtracking algorithm of Fig. 4.2 offers two possibilities of applying heuristics. One is in line 3 where the next unprocessed constraint can be chosen, the other is in line 5 where the next refinement can be chosen. These two choices influence the search space and the path through the search space. Good choices should increase efficiency of the backtracking algorithm.

Other fundamental reasoning problems are the *minimal label problem* CSP-MIN, the problem of finding the strongest entailed relation for each pair of variables from a given set of constraints, and the *entailment problem* CSPENT, i.e., decide whether a particular constraint is entailed by a set of constraints. As was shown for the corresponding temporal problems (see the next section), the entailment problem, the minimal label problem, and the consistency problem are equivalent under polynomial Turing reductions (Vilain et al., 1989; Golumbic and Shamir, 1993).

3. Spatial constraint calculi

In qualitative spatial reasoning it is common to consider a particular aspect of space such as topology, direction, or distance and to develop a system of qualitative relationships between spatial entities which cover this aspect of space to some degree and which appear to be useful from an applicational or from a cognitive perspective. If these relations are based on a set of jointly exhaustive and pairwise disjoint basic relations which is closed under several operations, it is possible to apply constraint based methods for reasoning over these relations (see Sec. 2). For this it is only necessary to give a composition table; either for all relations or for the basic relations plus a procedure for computing the compositions of complex relations.

The composition table should be obtained using the formal semantics of the relations. Otherwise it is not possible to verify correctness and completeness of the inferences. Formal semantics of the relations are also necessary for finding efficient reasoning algorithms which are essential for most applications. Without formal semantics it is sometimes not even possible to show that reasoning over a system of relations is decidable (e.g., the 9-intersection relations of Egenhofer, 1991, as shown by Grigni et al., 1995).

In the following subsections we survey some important approaches to the main aspects of space, topology, direction, and distance. Instead of summarising many different approaches, we focus on those approaches which have been formally analysed.

RCC-8 *Relation*	*Topological Constraints*
$\langle x, y\rangle \in \mathsf{DC}$	$x \cap y = \emptyset$
$\langle x, y\rangle \in \mathsf{EC}$	$i(x) \cap i(y) = \emptyset, x \cap y \neq \emptyset$
$\langle x, y\rangle \in \mathsf{PO}$	$i(x) \cap i(y) \neq \emptyset, x \not\subseteq y, y \not\subseteq x$
$\langle x, y\rangle \in \mathsf{TPP}$	$x \subset y, x \not\subseteq i(y)$
$\langle x, y\rangle \in \mathsf{TPP}^{-1}$	$y \subset x, y \not\subseteq i(x)$
$\langle x, y\rangle \in \mathsf{NTPP}$	$x \subset i(y)$
$\langle x, y\rangle \in \mathsf{NTPP}^{-1}$	$y \subset i(x)$
$\langle x, y\rangle \in \mathsf{EQ}$	$x = y$

Table 4.2. Topological interpretation of the eight base relations of RCC-8. All spatial regions are regular closed, i.e., $x = c(i(x))$ and $y = c(i(y))$. $i(\cdot)$ specifies the topological interior of a spatial region, $c(\cdot)$ the topological closure.

3.1 Topology

Topological distinctions between spatial entities are a fundamental aspect of spatial knowledge. Topological distinctions are inherently qualitative which makes them particularly interesting for qualitative spatial reasoning. Although there is a large body of work on topology developed in mathematics, this is not very well suited for qualitative spatial reasoning. Mathematical research on topology is not concerned with reasoning over topological relationships and as such does not provide us with any reasonable topological calculi and reasoning mechanisms (Gotts et al., 1996).

Topological approaches to qualitative spatial reasoning usually describe relationships between spatial regions rather than points, where spatial regions are subsets of some topological space. Most existing approaches on formalising topological properties of spatial regions are based on work from Whitehead, 1929, Clarke, 1981 and Clarke, 1985, who axiomatised mereotopology using a single primitive relation, the binary connectedness relation. Some approaches also distinguish between a mereological primitive, the parthood relation, and a topological primitive, the connected relation (Borgo et al., 1996). Using these primitive relations it is possible to define many other relations. A set of jointly exhaustive and pairwise disjoint relations which can be defined in all approaches of this kind are the eight relations $\mathsf{DC}, \mathsf{EC}, \mathsf{PO}, \mathsf{EQ}, \mathsf{TPP}, \mathsf{NTPP}, \mathsf{TPP}^{-1}, \mathsf{NTPP}^{-1}$. In the best known approach in this domain, the Region Connection Calculus by Randell et al., 1992, these relations are known as the RCC-8 relations. In Table 4.2 we defined the RCC-8 relations using the interior and exterior of spatial regions. Sample instances of the relations are given in Fig. 4.3. The relation symbols are abbreviations of their meanings: DisConnected, Externally Connected, Partially Overlapping, EQual, Tangential Proper Part, Non-Tangential Proper Part and the converse relations of the latter two relations.

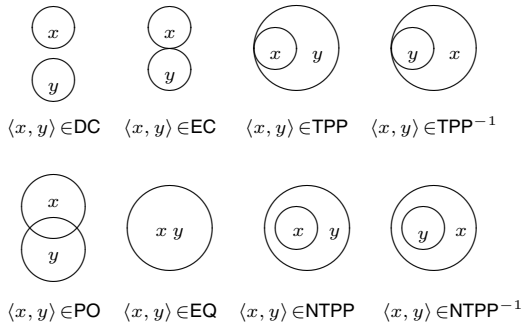


Figure 4.3. Two-dimensional examples for the eight basic relations of RCC-8.

What distinguishes the different approaches and what thereby influences the definable relations is the interpretation of the connectedness relation and the properties of the considered regions. Some approaches distinguish between open and closed regions (Randell and Cohn, 1989; Asher and Vieu, 1995) which allows, for instance, to define different kinds of contact. Asher and Vieu, 1995 distinguished between strong contact (two regions have points in common) and weak contact (two regions are disjoint but their topological closures share common points). Other approaches do not make this distinction and treat regions which are open, closed, or neither equally (Randell et al., 1992). The Region Connection Calculus (Randell et al., 1992) considers only the topological closure of regions: two regions are connected if their topological closures share a common point. Cohn et al., 1997 argue that this definition is more appropriate for commonsense spatial reasoning since there is "no reason to believe that some physical objects occupy closed regions and others open". Orthogonal to the interpretation of the connectedness relation is the distinction of what regions are considered. A very common restriction is to use only non-empty regular regions. As shown by Asher and Vieu, 1995, models based on Clarke's connectedness relation require all regions to be nonempty and regular. However, it is possible to specify additional properties of regions such as dimensionality, internal connectedness, i.e., whether a region consists of one-piece or of multiple pieces, or the existence of holes. The different approaches are compared in Cohn and Varzi, 1998 and Cohn and Varzi, 1999. In particular, the RCC-8 constraint language uses non-empty, regular closed regions which are subsets of a regular connected topological space. Regions do not have to be internally connected and are allowed to have holes.

All of these approaches have in common that the relations are axiomatised and defined in first-order logic which provides them with formal semantics. The formal properties of first-order theories based on a connectedness relation

were studied by Grzegorczyk, 1951, Dornheim, 1998, Pratt and Schoop, 1998 and Schoop, 1999. These are very expressive approaches and lead easily to undecidability of formal reasoning, i.e., logical implication in these theories is not decidable in general. When constraining oneself to less expressive languages such as constraint calculi, the computational costs are, of course, less. Constraint calculi can be regarded as the special case of first-order sentences where only existentially quantified region variables are used. Using the RCC-8 relations, we can state constraints such as, for example, $\mathsf{DC}(x_1, x_2)$, $\mathsf{TPP}^{-1}(x_2, x_3)$, $\{\mathsf{EC}, \mathsf{PO}\}(x_3, x_1)$. These are interpreted over the domain of regular closed regions of any regular topological space, such as, e.g., the n-dimensional Euclidean space. Now, given the composition table of RCC-8, we can use the algorithms introduced in the previous section in order to decide whether the set of constraint is consistent, or whether other constraints are logically implied. A prerequisite, however, is that a method for deciding consistency for some subset of the relation system formed by the the RCC-8 relations can be found. Bennett, 1994 gave an encoding of the RCC-8 relations in propositional modal logic and, thus proved that reasoning over the RCC-8 constraint language is decidable. In fact, this technique can be used as a decision method for RCC-8 constraint systems.

A different approach to defining topological relations was given by Egenhofer, 1991 in the area of spatial information systems. Egenhofer defined binary relations according to the 9 different intersections of the interior, exterior, and boundary of regions, hence, called *9-intersection*. Depending on the regions that are used, many different relations can be defined in this way (Egenhofer et al., 1994; Egenhofer and Franzosa, 1994). If only two-dimensional, internally connected regular regions without holes are considered and only emptiness or non-emptiness of the intersection is taken into account, this results in the same set of eight basic relations as definable in the above described approaches. In contrast to the connection based approaches, this approach is not provided with formal semantics which makes it very difficult to study its formal properties. For instance, attempts were made to identify sound and complete algorithms for reasoning over the eight relations defined by Smith and Park, 1992 and Egenhofer and Sharma, 1993, while it was taken for granted that path-consistency decides consistency if only basic relations are used. As shown by Grigni et al., 1995, this is not the case for Egenhofer's definition of the eight topological relations.

3.2 Orientation

Orientation is, like topology, very well suited for a qualitative approach. In everyday (non-technical) communication, orientation of spatial entities with respect to other spatial entities is usually given in terms of a qualitative category

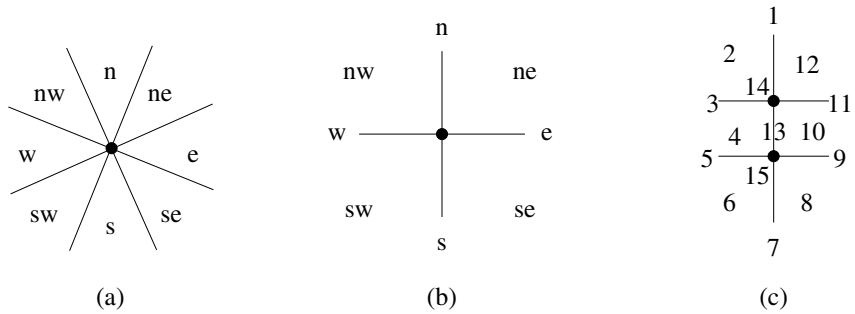


Figure 4.4. Orientation relations between points: (a) cone-based (b) projection-based (c) double-cross.

like "to the left of" or "northeast of" rather than using a numerical expression like "53 degrees" (which is certainly more common in technical communication like in aviation). Unlike the topological approaches we discussed in the previous section, orientation of spatial entities is a ternary relationship depending on the located object, the reference object, and the *frame of reference* which can be specified either by a third object or by a given direction. In the literature one distinguishes between three different kinds of frames of reference, extrinsic (external factors impose an orientation on the reference object), by some inherent property of the reference object), deictic (the orientation is imposed by the point of view from which the reference object is seen) (Hernàndez, 1994, p. 45). reference is given, orientation can be expressed in terms of binary relationships with respect to the given frame of reference.

Most approaches to qualitatively dealing with orientation are based on points as the basic spatial entities and consider only two-dimensional space. Frank, 1991 suggested different methods for describing the cardinal direction of a point with respect to a reference point in a geographic space, i.e., directions are in the form of "north", "east", "south", and "west" depending on the granularity. These are, however, just labels which can be equally termed as, for instance, "front", "right", "back", and "left" in a local space. Frank distinguishes between two different methods for determining the different sectors corresponding to the single directions: the *cone-based method* and the *projection-based method* (see Fig. 4.4). The projection-based approach allows us to represent the nine different relations (n, ne, e, se, s, sw, w, nw, eq) in terms of the point algebra by specifying a point algebraic relation for each of the two axes separately. This provides the projection-based approach (which is also called the *cardinal algebra*; see Ligozat, 1998) with formal semantics which were used by Ligozat, 1998 to study its computational properties. In particular, Ligozat found that reasoning with the cardinal algebra is NP-complete (See below in Sec. 4) and,

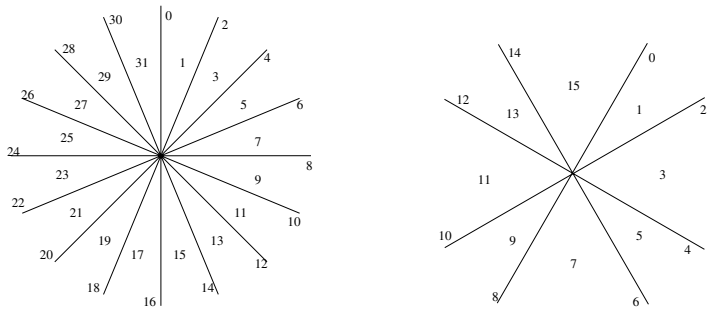


Figure 4.5. Two different Star calculi, one with 8 lines forming 33 relations and one with 4 lines forming 17 relations.

further, identified a maximal tractable subset of the cardinal algebra by using the concept of *preconvex* relations, a method which has already been used for Allen's interval algebra (Ligozat, 1996).

A generalisation of these calculi was proposed and analysed by Renz and Mitra, 2004. Their calculus, the *Star calculus* (see Fig. 4.5), is based on a number of n lines l_i with given angles δ_i (for arbitrary n) which define $2n$ sectors and $4n + 1$ basic relations. The number of lines and the angles of the sectors can be adopted to the given application, so the Star calculus can be used for representing and reasoning about qualitative directions of arbitrary granularity. The Star calculus has some interesting properties. For example, it can be shown that when having three or more lines it is possible to emulate a coordinate system which is due to having the lines as separate basic relations. This removes the distinction made between qualitative and quantitative representation and also means that qualitative reasoning methods like path-consistency cannot be complete for deciding consistency. Renz and Mitra therefore proposed to combine the lines and the sectors and to consider them as new basic relations. In both cases the full calculus is NP-hard while reasoning over the basic relations is tractable.

A further point-based approach was developed by Freksa, 1992, the so-called *double-cross* calculus, which defines the direction of a located point to a reference point with respect to a perspective point. Within this approach three axes are used: one is specified by the perspective point and the reference point, the other two axes are orthogonal to the first one and are specified by the reference point and the perspective point. These axes define 15 different ternary basic relations (see Fig. 4.4c). The computational properties of this calculus have been studied by Scivos and Nebel, 2001. It turned out that the consistency problem is NP-hard even if only the 15 basic relations are used.

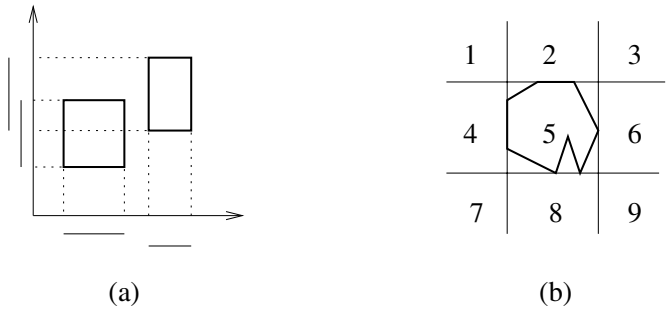


Figure 4.6. Orientation relations between extended entities: (a) rectangle algebra (b) direction-relation matrix.

Developing orientation relations between extended spatial entities is much more difficult than between points. Extended objects often have their own intrinsic directions like a natural front. Also, the direction between extended objects with complex shapes such that, for instance, their convex hulls intersect is not at all clear. Even for simpler objects there is often no agreement about which natural language expression describes their orientational relationship best. Therefore, most approaches use approximations to the spatial regions or use only spatial regions of a particular kind. One approach which has been chosen by many researchers results in restricting all regions to be rectangles whose sides are parallel to the axes determined by the frame of reference (Guesgen, 1989; Papadias and Theodoridis, 1997; Balbiani et al., 1998). In this approach all regions can be represented by their projections to the defining axes which corresponds to having Allen's interval algebra (Allen, 1983) for each axis separately, i.e., every relation is a pair of two interval relations (see Fig. 4.6a). For two-dimensional space this results in 13×13 different basic relations (also called the *rectangle algebra*; see Balbiani et al., 1998) whose formal semantics are provided by the interval algebra.

Balbiani et al., 1998 and Balbiani et al., 1999a studied the formal properties of the rectangle algebra (and also of the *block algebra* which is the n-dimensional extension of the interval algebra; see Balbiani et al., 1999b). NP-completeness of the algebra carries over from the interval algebra. Balbiani et al., 1998 and Balbiani et al., 1999a identified a tractable subset of the rectangle algebra following a line of reasoning which has been introduced by Ligozat, 1996, namely, by considering convex and preconvex relations. Unlike for the interval algebra, the set of preconvex relations is not closed under the fundamental operations. Thus, Balbiani et al. extended the concept of preconvexity by distinguishing between weakly and strongly preconvex relations, and show that the set of strongly preconvex relations is a tractable subset of the rectangle

algebra for which path-consistency is sufficient for deciding consistency. In fact the rectangle algebra represents more than just orientation between two rectangles but also their topological relations. Hence, the rectangle algebra can be regarded as an approach to combining topology and orientation. However, because of the large number of relations of the rectangle algebra (a total number of 2^{169} relations) reasoning even over a tractable subset can be very inefficient.

An interesting but less expressive approach to representing orientational relationships between extended spatial entities was introduced by Goyal Egenhofer, 2001. Their calculus consists of a 3×3 *direction-relation matrix* which represents the 9 sectors formed by the minimal bounding axes of an extended spatial entity (see Fig. 4.6b). For each sector it is possible to specify whether the located object is contained in the sector or not, or (non-qualitatively) to which degree the located object is contained in the sector.

Skiadopoulos and Koubarakis, 2005 developed reasoning algorithms for this calculus and analysed its computational properties, but their algorithms are not based on constraint based methods like path-consistency.

3.3 Distance

Together with topology and orientation, distance is one of the most important aspects of space. Unlike the other two, distance is a scalar entity. Dealing with distance information is an important cognitive ability in our everyday life. In order to grab something, for instance, we must be good in judging distances. When communicating about distances, we usually use qualitative categories like "A is close to B" or qualitative distance comparatives like "A is closer to B than to C", but also numerical values like "A is about one meter away from B". As indicated by the above examples, one can distinguish between absolute distance relations (the distance between two spatial entities) and relative distance relations (the distance between two spatial entities as compared to the distance to a third entity). While absolute distance can be represented either qualitatively or quantitatively, relative distance is purely qualitative. When representing absolute distance in a qualitative way, this also depends on the scale of space which is used. Montello, 1993 suggests four different kinds of scales of space: figural space, vista space, environmental space, and geographic space.

Most approaches to qualitative distance consider points as the basic entities. Absolute distance relations are obtained, e.g., by dividing the real line into several sectors such as "very close", "close", "commensurate", "far", and "very far" depending on the chosen level of granularity (Hernàndez et al., 1995). Relative distance can be obtained by comparing the distance to a given reference distance which results in ternary relations such as "closer than", "equidistant", or "farther than". Reasoning about qualitative distances leads to several

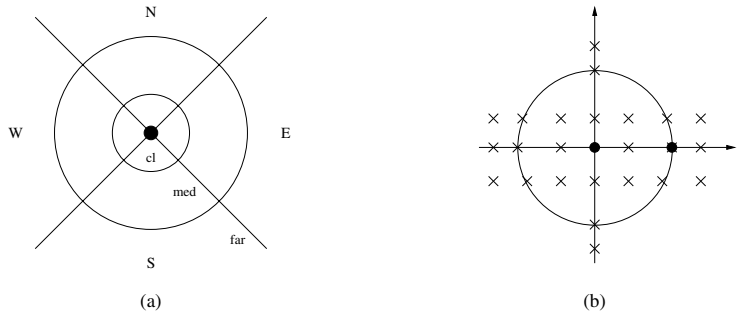


Figure 4.7. Different approaches to representing positional information: (a) absolute distances combined with cone-based orientation (Clementini et al., 1997) (b) relative distances combined with projection-based orientation (Isli and Moratz, 1999).

difficulties. For instance, given a sequence of collinear points $p_1, \ldots, p_n$ such that p_i is close to p_{i+1} for every i, for which n is p_n far from p_1? Moreover, combining distance relations does not only depend on the distances itself but also on the position of the corresponding points. For instance if point B is far from A and C is far from B, then C can be very far from A if A, B, and C are aligned and if B is between A and C; or C can be close to A if the angle between AB and BC is small. Therefore, it seems advisable to study distance in combination with orientation. This combination is called *positional* information.

One approach for developing a position calculus is by Clementini et al., 1997, who combine a cone-based orientation approach with absolute distance relations (see Fig. 4.7a). Clementini et al. present different procedures for computing the composition of two positional relations (A, B) and (B, C). They consider three special cases where BC is the same, opposite, or orthogonal direction to AB. Another approach is by Isli and Moratz, 1999, who propose several position calculi on various levels of granularity by combining relative distance relations with different approaches to orientation such as the projection-based approach (see Fig. 4.7b) or the double-cross calculus. The computational properties of these approaches have not been studied yet.

4. Computational complexity

Since we are interested in automated reasoning with the spatial calculi described above, it is a good idea to get an understanding of how computationally demanding reasoning in these calculi is. Here computational complexity theory is the right theoretical tool.

In the field of computational complexity (Papadimitriou, 1994), computational problems are classified according to their need for resources for solving

them, usually the running time and the memory consumption. This allows to compare the complexity of different problems and to design algorithms for a whole class of problems. For classifying computational problems, they are usually expressed as *decision problems*, i.e., problems that require a simple yes/no answer. Such problems can be equivalently viewed as formal languages over some alphabet Σ, which is formed by all yes-instances.

Most problems can be easily translated into an equivalent decision problem. Assume for example the problem of finding a satisfying truth assignment for a propositional formula. The corresponding decision problem is the problem SAT: given a set of variables V and a propositional formula ϕ over V in CNF, is there a satisfying truth assignment for ϕ? The complexity of a decision problem is usually measured according to the worst-case running time or memory consumption of the best possible algorithm. If we now can prove lower bounds on the runtime for the decision problem then these lower bounds apply obviously to the original problem as well.

Running time as well as memory consumption of an algorithm depends on the size n of its input, i.e., on the size of the problem instance, and can be expressed as a function $f(n)$. For classifying algorithms according to their running time, the asymptotical behaviour is more important than f itself. This is specified in terms of the *O-notation* which gives an upper bound on the running time within a constant factor (Cormen et al., 1990). An algorithm with a running time of $O(n^3)$ or faster is usually considered to be efficient. In areas like database systems where instances have a very large size, a running time of $O(n^3)$ is too slow. In these areas efficient algorithms should have a linear running time.

4.1 Tractability and NP-completeness

There is a large number of different complexity classes that are used to categorise decision problems (Johnson, 1990). Particularly important is the class of decision problems that can be solved in polynomial time using a deterministic algorithm. This complexity class is called P and it is considered to be the class of efficiently solvable problems. Problems in P are also called *tractable* problems, problems outside *P* are called *intractable* problems.

Interestingly, there exists a large class of problems for which nobody has found polynomial-time algorithms yet, but it appears equally hard to prove that no such algorithms exist. In order to capture these problems, one extends the notion of algorithm. The class of problems solvable in polynomial time using a non-deterministic algorithm is called NP, which is equivalent to specifying that a given solution of an NP problem can be verified in polynomial time using a deterministic algorithm. In the sequel, an algorithm will always be a deterministic algorithm, unless otherwise stated. It is clear that P is a subset

of NP, but it is not known whether P is a proper subset of NP or whether P is equal to NP, which is called the $\mathsf{P} \stackrel{?}{=} \mathsf{NP}$ problem.

An important method of comparing problems is specifying a *reduction* from one problem to another. Given two problems $A, B \subseteq \Sigma^*$, problem A can be *reduced* to problem B by giving a constructive transformation $f\colon \Sigma^* \to \Sigma^*$ such that $f(x) \in B$ if and only if $x \in A$. If f can be computed in polynomial time, the reduction is a *polynomial (time) reduction*. If A is polynomially reducible to B (written as $A \leq_p B$), then any polynomial time algorithm for solving B can be used to solve A. Thus, for showing that a particular decision problem A is in P, it is sufficient to find another problem $B \in \mathsf{P}$ such that $A \leq_p B$.

A decision problem A is said to be NP-*hard* if any other problem in NP can be polynomially reduced to A. An NP-hard problem which is itself contained in NP is called NP-*complete*. NP-complete problems are the most difficult problems in NP. In fact, most of the problems for which nobody has found an efficient algorithms yet but which are resistant against proving them to be intractable fall into this class. In order to prove a decision problem A to be NP-hard, it is sufficient to find another NP-hard problem that can be polynomially reduced to A. The first problem that was identified to be NP-complete is the SAT problem (Cook, 1971). In this work we use the following NP-complete propositional decision problems (Garey and Johnson, 1979):

Given: A set of variables V and a propositional formula ϕ over V in CNF such that each clause of ϕ has exactly three literals.

Questions:

1. Is there a satisfying truth assignment for ϕ? (3SAT)
2. Is there a satisfying truth assignment for ϕ such that each clause has at least one true literal and at least one false literal? (NOT-ALL-EQUAL-3SAT)
3. Is there a satisfying truth assignment for ϕ such that each clause has exactly one true literal? (ONE-IN-THREE-3SAT)

Some variants of the propositional satisfiability problem are solvable in polynomial time. This includes the 2SAT problem, the propositional satisfiability problem of Krom formulae, and the HORNSAT problem, the propositional satisfiability problem of Horn formulae, which is of particular importance in this work. It is generally believed that $\mathsf{P} \neq \mathsf{NP}$, and, hence, that NP-complete problems are intractable. This is also the assumption of this work. So far, any algorithm for an NP-complete problem has at least super-polynomial running time.

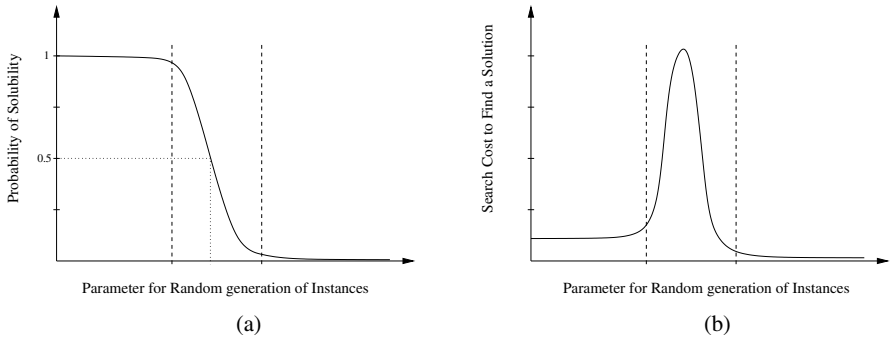


Figure 4.8. Typical phase-transition behaviour of randomly generated instances.

4.2 Phase Transitions

Having proved a problem to be NP-complete is not the end of the computational analysis of a problem, but rather its beginning. NP-completeness is just a worst-case measure of a problem. It means that for any algorithm there exist instances which cannot be solved in polynomial time. It is possible that only one in a million instances is very hard and that the other instances can be solved easily.

There are several ways to deal with NP-complete problems. One way is to develop efficient approximation algorithms which are correct but not complete for deciding either solubility or insolubility. Another way is to use complete algorithms which require exponential time in the worst-case and to develop heuristics which solve many instances efficiently. In all cases the effectiveness of new algorithms and heuristics should be verified using a large number of instances. Since it is usually not easy to obtain a large number of real-world instances, many researchers generate instances randomly with respect to different control parameters.

Cheeseman et al., 1991 found that randomly generated instances of the NP-complete problems they studied had a very special behaviour: when ordering these instances according to a particular problem-dependent parameter, there are three different regions with respect to the solubility of the instances that occur when changing the parameter. In one region instances are soluble with a very high probability, in one region instances are insoluble with a very high probability, and in between these two regions there is a very small region where the probability of solubility of these instances changes abruptly from very high to very low (see Fig. 4.8(a)). Cheeseman et al., 1991 called this region the

phase-transition. In this region a small change of the local parameter leads to a large change in the solubility of the instances.

Cheeseman et al., 1991 further found that almost all hard instances are located in the phase-transition region. In general, instances in the phase transition appear to be harder than instances in the other two regions (see Fig. 4.8(b)). This is because instances in soluble region are under-constrained and for this reason any search method finds a solution very fast without much backtracking. Similarly, instances in the insoluble region are over-constrained and for this reason search methods fail quite early when searching through the space of possible solutions. In some studies, however, it turned out that some under-constrained instances are particularly hard (Gent and Walsh, 1996).

The behaviour of randomly generated instances of NP-complete problems described by Cheeseman et al. was found by many researchers for many NP-complete problems, although satisfiability problems were the most studied problems. A typical parameter for satisfiability problems that causes a phase transition is the ratio of clauses-to-variables. An interesting selection of papers on the topic can be found in Hogg et al., 1996.

4.3 How to prove NP-hardness and NP membership for spatial CSPs

In order to prove NP-hardness of a decision problem, in our case the consistency problem of a set of spatial constraints CSPSAT, it is sufficient to find another NP-hard problem that can be polynomially reduced to the problem at hand. Usually this has to be done in a different way for each new problem again and again and it is in many cases a difficult task to find a new transformation and to prove that it is a one to one transformation. The difficulty of this problem can be estimated when considering that new NP-hardness proofs often deserve a publication.

When looking at spatial CSPs over different sets of relations it is striking that they all have the same structure with different relations. One might expect that the same reduction with different parameters can be used for different sets of relations and that a general transformation scheme can be used. In this section we present a scheme which we developed for proving NP-hardness of different subsets of RCC-8 and which seems to be general in the way that the parameters of the scheme can be found by exhaustive search over possible relations, no matter what the relations are. So the transformations could essentially be identified automatically for any system of relations.

Our scheme uses a transformation from a propositional satisfiability problem to CSPSAT($\mathcal{S}$) where $\mathcal{S}$ is a subset of a system of relations $2^{\mathcal{B}}$ by constructing a set of spatial constraints Θ for every instance $\mathcal{I}$ of the propositional satisfiability problem, such that Θ is consistent if and only if $\mathcal{I}$ is a positive instance. The

propositional satisfiability problems we use are 3SAT, the problem of deciding whether there is a truth assignment for a set of clauses where each clause has exactly three literals, as well as two variants of 3SAT where truth assignments of particular types are required. These variants are NOT-ALL-EQUAL-3SAT, the problem of deciding whether there is a truth assignment such that for every clause at least one literal is assigned $true$ and one literal is assigned $false$, and ONE-IN-THREE-3SAT, the problem of deciding whether there is a truth assignment such that for every clause exactly one literal in every clause is assigned $true$. All three decision problems are NP-hard (Schaefer, 1978).

The different transformations have in common that every variable v of the propositional satisfiability problem is transformed to two constraints $x_v\{R_t, R_f\}$ y_v and $x_{\neg v}\{R_t, R_f\}y_{\neg v}$ corresponding to the positive and the negative literal of v, where R_t and R_f are relations of $\mathcal{S}$ with $R_t \cap R_f = \emptyset$. v is assigned $true$ if and only if $x_v\{R_t\}y_v$ holds and assigned $false$ if and only if $x_v\{R_f\}y_v$ holds. Since the two literals corresponding to a variable need to have opposite assignments, we have to make sure that $x_v\{R_t\}y_v$ holds if and only if $x_{\neg v}\{R_f\}y_{\neg v}$ holds, and *vice versa*, for which additional *polarity constraints* are required. In addition, every literal occurrence l of the propositional satisfiability problem is transformed to the constraint $x_l\{R_t, R_f\}y_l$, where $x_l\{R_t\}y_l$ holds if and only if l is assigned $true$. In order to assure the correct assignment of positive and negative literal occurrences with respect to the corresponding variable, polarity constraints are required again. For instance, if the variable v is assigned $true$, i.e., $x_v\{R_t\}y_v$ holds, then $x_p\{R_t\}y_p$ must hold for every positive literal occurrence p of v, and $x_n\{R_f\}y_n$ must hold for every negative literal occurrence n of v. Further, *clause constraints* have to be added to ensure that the clause requirements of the specific propositional satisfiability problem are satisfied. For example, if $\{i, j, k\}$ is a clause of an instance of ONE-IN-THREE-3SAT, then exactly one of the constraints $x_i\{R_t\}y_i$, $x_j\{R_t\}y_j$, and $x_k\{R_t\}y_k$ must hold.

According to this scheme, all we have to do in order to find a transformation is to identify relations $R_f, R_t \in \mathcal{S}$, the polarity constraints which enable to propagate the assignment of literal occurrences to other literal occurrences, and the clause constraints which ensure that properties of clauses also hold for their transformation. These constraints can be found by exhaustively assigning and testing the polarity CSP of figure 4.9(a) and based on this, the clause CSP of figure 4.9(b). If it is possible to identify the polarity and the clause constraints, then we have found a polynomial transformation from a propositional satisfiability problem to the consistency problem of $\mathcal{S}$.

The next step is to show that this transformation is a many-to-one transformation such that whenever we have a positive instance of the propositional satisfiability problem we get a positive instance of the consistency problem. Unlike finding polarity and clause constraints, this part of the NP-hardness proof

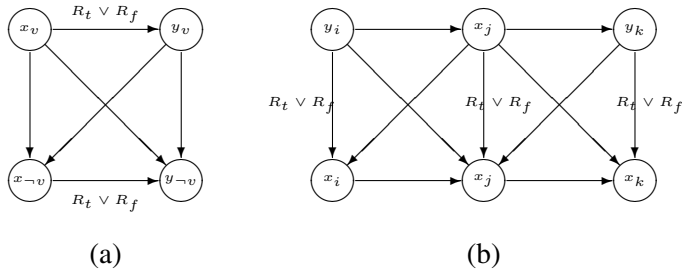


Figure 4.9. The polarity constraints (a) ensure that positive and negative literals of the same variable have opposite assignments. The clause constraints (b) ensure that the clause requirements of the particular 3SAT problem are satisfied.

cannot be automated as it depends on the domains we are using. However, since the CSP we get is very structured with only the polarity constraints and the clause constraints, this can be easy to show in many cases. It is actually an advantage that the domain we are using is infinite as it allows to treat the different polarity and clause constraints almost independently. Examples for transforming propositional satisfiability problems to CSPSAT($\mathcal{S}$) for different subsets $\mathcal{S}$ of RCC-8 can be found in Renz and Nebel, 1999.

The next step in the complexity analysis of a given spatial calculus is to prove NP membership of CSPSAT. Recall that in order to show that a decision problem is a member of NP, we have to show that a possible solution can be checked in polynomial time. So for showing that CSPSAT is in NP for a system of relations $2^{\mathcal{B}}$ over a domain $\mathcal{D}$ it is sufficient to show that CSPSAT($\mathcal{B}$) over $\mathcal{D}$ is a member of NP. While for many NP-complete problems the NP-membership proof is easier than showing NP-hardness, it is the other way around for spatial CSPs. This is due to the fact that we might have to check arbitrary spatial entities which might not even be representable in a computational framework (see e.g., Renz, 1998) and due to the infinity of the domain $\mathcal{D}$. Proving NP membership can be very difficult and has to be done for each system of relations and for each domain separately. Consider for example the RCC-8 relations and Egenhofer's relations. The composition table of the relations are the same, but the domains are different. While the RCC-8 domain consists of regular subsets of a topological space, Egenhofer's domains consist of connected two-dimensional regions without holes which is much more restricted. The consequence of this is that while RCC-8 is in NP (Renz, 1998), NP membership of Egenhofer's calculus is still an open problem (Grigni et al., 1995). Instead of proving NP membership by showing that a given solution can be verified in polynomial time, we can also give a polynomial time decision procedure for CSPSAT($\mathcal{B}$) over $\mathcal{D}$ and show that whenever the decision procedure recognises an instance

Θ as consistent, there is an instantiation of all variables in Θ with values of the domain $\mathcal{D}$ such that all constraints of Θ are satisfied. Having a polynomial decision procedure is a stronger result and implies NP membership. In the next session we will look at how such decision procedures can be identified.

5. Identifying tractable subsets of spatial CSPs

Reasoning about most interesting spatial calculi is NP-hard. This, however, is often true only for the full calculus, i.e., if all relations $2^{\mathcal{B}}$ can be used. If we restrict ourselves to subsets $\mathcal{S} \subseteq 2^{\mathcal{B}}$ of the full set of relations, it might be possible that reasoning over this subset is tractable. Ideally we are interested in finding maximal tractable subsets of $2^{\mathcal{B}}$ which are those subsets which are tractable and which become NP-hard if any other relation is added. This represents the boundary between tractability and NP-hardness. Some subsets are obviously tractable such as the set of relations that contain the identity relations as a disjunct. The subsets that are most interesting are those that contain all the basic relations $\mathcal{B}$. So as a minimal requirement and as a first step we have to show that the set of basic relations is tractable.

For RCC-8, Renz and Nebel, 1999 showed tractability of the basic relations by developing a polynomial transformation of RCC-8 constraints into SAT formulae. Those RCC-8 relations that transform into a Horn formula together form a tractable subset. Altogether 64 relations were identified in this way, among them were all the basic relations. While we could try to develop a new algorithm or a new transformation for every spatial calculus and for different subsets of them, it is highly desirable that the path-consistency algorithm (see Sec. 2.1) can be used for deciding consistency of tractable subsets. If this is the case then consistency can be decided purely by algebraic operations on the relations without having to fall back to the infinite domains. And we have to deal with the domains only once for proving that path-consistency decides consistency. Obviously, this again depends strongly on the domains and the relations that are used and cannot be generalised. Therefore we have to find a new tractability proof for every set of basic relations over every domain. This can be very complicated as we have to deal with infinite domains.

For RCC-8, for example, the proof that the path-consistency algorithm decides consistency for the basic relations (actually for a larger set of relations) was done as follows (Renz and Nebel, 1999). First, it was analysed how applying path-consistency can lead to an inconsistency. Then it was shown that whenever the path-consistency algorithm detects an inconsistency, positive unit resolution applied to the SAT encoding of RCC-8 produces the empty clause. In Renz, 1998 an algorithm was presented which computes an instantiation for all variables of any consistent set of constraints over the RCC-8 basic relations.

This algorithm works for Euclidean spaces in all dimensions d. For $d \geq 3$ it also works for connected regions without holes.

Once it has been shown that path-consistency decides consistency for the basic relations, it is possible to try to extend the set of relations and to identify larger tractable subsets. There are basically two general methods which can be used for extending tractability of subsets of relations to larger subsets: the closure method (Renz and Nebel, 1999) and the refinement method (Renz, 1999). We will describe these methods in the following section. In particular the refinement method seems very powerful and will be presented in more detail.

5.1 Closure of sets of relations

Given a system of relations $2^{\mathcal{B}}$, the number of subsets $\mathcal{S} \subseteq 2^{\mathcal{B}}$ that we might have to analyse for a computational analysis is huge, namely, $2^{(2^{|\mathcal{B}|})}$. This number can be slightly reduced if only those subsets are considered that contain all the basic relations and possibly the universal relation. Fortunately, we can reduce the number of subsets further by noting that the computational complexity associated with an arbitrary subset $\mathcal{S}$ is identical to the complexity associated with the closure of this subset under composition, intersection, and converse, denoted by $\widehat{\mathcal{S}}$—an observation that was first used in determining a maximal tractable subset of Allen's interval calculus (Nebel and Bürckert, 1995, Theorem 14). Renz and Nebel, 1999 proved this for arbitrary systems of relations and came up with the following theorem.

THEOREM 4.1 *Let $\mathcal{C}$ be a set of binary relations that is closed under composition, intersection, and converse. Then for any subset $\mathcal{S} \subseteq \mathcal{C}$ that contains the universal relation, the problem* $\mathsf{CSPSAT}(\widehat{\mathcal{S}})$ *can be polynomially reduced to* $\mathsf{CSPSAT}(\mathcal{S})$.

Note that Theorem 4.1 holds only if there exists an infinite supply of fresh variables; this is not always the case (e.g., bounded variable problems which are studied in logic and model theory). Another requirement of Theorem 4.1 is the possibility to specify more than one constraint for each pair of variables. Otherwise the identity relation must be contained in $\mathcal{S}$. The following corollary specifies how Theorem 4.1 will be used.

COROLLARY 4.2 *Let $\mathcal{S}$ be a subset of $2^{\mathcal{B}}$.*

1 $\mathsf{CSPSAT}(\widehat{\mathcal{S}}) \in \mathsf{P}$ *if and only if* $\mathsf{CSPSAT}(\mathcal{S}) \in \mathsf{P}$.

2 $\mathsf{CSPSAT}(\mathcal{S})$ *is* NP*-hard if and only if* $\mathsf{CSPSAT}(\widehat{\mathcal{S}})$ *is* NP*-hard.*

The first statement of Corollary 4.2 can be used to increase the number of elements of tractable subsets of CSPSAT considerably. With the second statement of Corollary 4.2 NP-hardness proofs of CSPSAT can be used to

exclude certain relations from being in any tractable subset of CSPSAT. In any case, we will have to analyse only those subsets that are closed under composition, converse and intersection.

The computational analysis of RCC-8 shows how powerful this method is. The closure of the set of 64 relations that transform to Horn formulae consists of 148 relations (called $\widehat{\mathcal{H}}_8$) and turns out to be a maximal tractable subset of RCC-8. Furthermore, it has been shown (Renz and Nebel, 1999) that path-consistency decides CSPSAT($\widehat{\mathcal{H}}_8$).

5.2 The refinement method

In this subsection we present a general method for proving tractability of reasoning over disjunctions of a JEPD set $\mathcal{B}$ of binary relations over a domain $\mathcal{D}$ which are atoms of a relation algebra, i.e., a method for proving tractability of CSPSAT($\mathcal{S}$) for sets $\mathcal{S} \subseteq 2^{\mathcal{B}}$ (see Sec. 2.2). In order to do so, this method requires a subset $\mathcal{T}$ of $2^{\mathcal{B}}$ for which path-consistency is already known to decide CSPSAT($\mathcal{T}$). Then the method checks whether it is possible to refine every constraint involving a relation in $\mathcal{S}$ according to a particular refinement scheme to a constraint involving a relation in $\mathcal{T}$ without changing consistency. The following definition will be central for this method.

DEFINITION 4.3 (REDUCTION BY REFINEMENT)
Let $\mathcal{S}, \mathcal{T} \subseteq 2^{\mathcal{B}}$. $\mathcal{S}$ can be reduced by refinement *to $\mathcal{T}$, if the following two conditions are satisfied:*

1. *for every relation $S \in \mathcal{S}$ there is a relation $T_S \in \mathcal{T}$ with $T_S \subseteq S$,*
2. *every path-consistent set Θ of constraints over $\mathcal{S}$ can be refined to a set Θ' of constraints over $\mathcal{T}$ by replacing $x_i S x_j \in \Theta$ with $x_i T_S x_j \in \Theta'$ for $i < j$, such that enforcing path-consistency to Θ' does not result in an inconsistency.*

Note that in the above definition constraints $x_i S x_j$ are refined only for $i < j$. This is no restriction, as by enforcing path-consistency the converse constraint $x_j S^{\smile} x_i$ will also be refined. Rather it offers the possibility of refining, e.g., converse relations to other than converse sub-relations, i.e., if, for instance, R is refined to r, $R^{\smile}$ can be refined to a relation other than $r^{\smile}$. This property of a set of relations can be used to derive its tractability.

LEMMA 4.4 *If path-consistency decides* CSPSAT($\mathcal{T}$) *for a set $\mathcal{T} \subseteq 2^{\mathcal{B}}$, and $\mathcal{S}$ can be reduced by refinement to $\mathcal{T}$, then path-consistency decides* CSPSAT($\mathcal{S}$).

Proof. Let Θ be a path-consistent set of constraints over $\mathcal{S}$. Since $\mathcal{S}$ can be reduced by refinement to $\mathcal{T}$, there is by definition a set Θ' of constraints over $\mathcal{T}$ which is a refinement of Θ such that enforcing path-consistency to Θ' does

not result in an inconsistency. Path-consistency decides $\mathsf{CSPSAT}(\mathcal{T})$, so Θ' is consistent, and, hence, Θ is also consistent. ■

Since path-consistency can be enforced in cubic time, it is sufficient for proving tractability of $\mathsf{CSPSAT}(\mathcal{S})$ to show that $\mathcal{S}$ can be reduced by refinement to a set $\mathcal{T}$ for which path-consistency decides $\mathsf{CSPSAT}(\mathcal{T})$. Note that for refining a constraint xSy ($S \in \mathcal{S}$) to a constraint xT_Sy ($T_S \in \mathcal{T}$), it is not required that T_S is also contained in $\mathcal{S}$. Thus, with respect to common relations the two sets $\mathcal{S}$ and $\mathcal{T}$ are independent of each other. This is in contrast to Theorem 4.1 which states that the tractability of a set of relations implies the tractability of its closure.

We will now present a method for showing that a set of relations $\mathcal{S} \subseteq 2^{\mathcal{B}}$ can be reduced by refinement to another set $\mathcal{T} \subseteq 2^{\mathcal{B}}$. In order to manage the different refinements, a *refinement matrix* is introduced that contains for every relation $S \in \mathcal{S}$ all specified refinements.

DEFINITION 4.5 (REFINEMENT MATRIX)
A refinement matrix M *of* $\mathcal{S}$ *has* $|\mathcal{S}| \times 2^{|\mathcal{B}|}$ *Boolean entries such that for* $S \in \mathcal{S}$, $R \in 2^{\mathcal{B}}$, $M[S][R] = true$ *only if* $R \subseteq S$.

For example, if we want to build a refinement matrix which states that that the relation $\{\mathsf{DC}, \mathsf{EC}, \mathsf{PO}, \mathsf{TPP}\}$ can be refined to the relations $\{\mathsf{DC}, \mathsf{TPP}\}$ and $\{\mathsf{DC}\}$, then we set $M[\{\mathsf{DC}, \mathsf{EC}, \mathsf{PO}, \mathsf{TPP}\}][R]$ is $true$ only for $R = \{\mathsf{DC}, \mathsf{TPP}\}$ and for $R = \{\mathsf{DC}\}$ and $false$ for all other relations $R \in 2^{\mathcal{B}}$. M is called the *basic refinement matrix* if $M[S][R] = true$ if and only if $S = R$.

Renz, 1999 proposes the algorithm CHECK-REFINEMENTS (see Fig. 4.10) which takes as input a set of relations $\mathcal{S}$ and a refinement matrix M of $\mathcal{S}$. The algorithm uses triples of relations $T = (R_{12}, R_{23}, R_{13})$ which represent sets of constraints $\{xR_{12}y, yR_{23}z, xR_{13}z\}$ for some variables x, y, z. It computes all possible path-consistent triples of relations R_{12}, R_{23}, R_{13} of $\mathcal{S}$ (step 4), and enforces path-consistency (using a standard procedure PATH-CONSISTENCY) to every refinement $R'_{12}, R'_{23}, R'_{13}$ for which $M[R_{ij}][R'_{ij}] = true$ for all $i, j \in \{1, 2, 3\}, i < j$ (steps 5,6). If one of these refinements results in the empty relation, the algorithm returns `fail` (step 7). Otherwise, the resulting relations $R''_{12}, R''_{23}, R''_{13}$ are added to M by setting $M[R_{ij}][R''_{ij}] = true$ for all $i, j \in \{1, 2, 3\}, i < j$ (step 8). This is repeated until M has reached a fixed point (step 9), i.e., enforcing path-consistency on any possible refinement does not result in new relations anymore. If no inconsistency is detected in this process, the algorithm returns `succeed`.

A similar algorithm, GET-REFINEMENTS, returns the revised refinement matrix if CHECK-REFINEMENTS returns `succeed` and the basic refinement matrix if CHECK-REFINEMENTS returns `fail`. Since $\mathcal{B}$ is a finite set of relations, M can be changed only a finite number of times, so both algorithms

Algorithm: CHECK-REFINEMENTS
Input: A set $\mathcal{S}$ and a refinement matrix M of $\mathcal{S}$.
Output: `fail` if the refinements specified in M can make a path-consistent triple of constraints over $\mathcal{S}$ inconsistent; `succeed` otherwise.

1. changes $\leftarrow true$
2. *while* changes *do*
3. $\quad oldM \leftarrow M$
4. $\quad$ *for every* path-consistent triple $T = (R_{12}, R_{23}, R_{13})$ of relations over $\mathcal{S}$ *do*
5. $\quad\quad$ *for every* refinement $T' = (R'_{12}, R'_{23}, R'_{13})$ of T with $oldM[R_{12}][R'_{12}] = oldM[R_{23}][R'_{23}] = oldM[R_{13}][R'_{13}] = true$ *do*
6. $\quad\quad\quad T'' \leftarrow$ PATH-CONSISTENCY(T')
7. $\quad\quad\quad$ *if* $T'' = (R''_{12}, R''_{23}, R''_{13})$ contains the empty relation *then return* `fail`
8. $\quad\quad\quad$ *else do* $M[R_{12}][R''_{12}] \leftarrow true$, $M[R_{23}][R''_{23}] \leftarrow true$, $M[R_{13}][R''_{13}] \leftarrow true$
9. $\quad$ *if* $M = oldM$ *then* changes $\leftarrow false$
10. *return* `succeed`

Figure 4.10. Algorithm CHECK-REFINEMENTS.

always terminate. If $n = |2^{\mathcal{A}}|$ is the total number of relations, then there are at most n^3 possible triples of relations in step 4, at most n^3 possible refinements of each triple in step 5, and at most n^2 iterations of the *while* loop. Thus, a rough estimation of the worst-case running time of both algorithms leads to $O(n^8)$.

LEMMA 4.6 *Let Θ be a path-consistent set of constraints over $\mathcal{S}$ and M a refinement matrix of $\mathcal{S}$. For every refinement Θ' of Θ with $x_i R' x_j \in \Theta'$ only if $x_i R x_j \in \Theta$, $i < j$, and $M[R][R'] = true$, the following holds: if* CHECK-REFINEMENTS$(\mathcal{S}, M)$ *returns* `succeed`*, enforcing path-consistency to Θ' does not result in an inconsistency.*

If CHECK-REFINEMENTS returns `succeed` and GET-REFINEMENTS returns M', we have pre-computed all possible refinements of every path-consistent triple of variables as given in the refinement matrix M'. Thus, applying these refinements to a path-consistent set of constraints can never result in an inconsistency when enforcing path-consistency.

THEOREM 4.7 *Let $\mathcal{S}, \mathcal{T} \subseteq 2^{\mathcal{B}}$, and let M be a refinement matrix of $\mathcal{S}$. Let M' be the refinement matrix returned by* GET-REFINEMENTS$(\mathcal{S}, M)$. *If for every $S \in \mathcal{S}$ there is a $T_S \in \mathcal{T}$ with $M'[S][T_S] = true$, then $\mathcal{S}$ can be reduced by refinement to $\mathcal{T}$.*

By Lemma 4.4 and Theorem 4.7 we have that the procedures CHECK-REFINEMENTS and GET-REFINEMENTS can be used to prove tractability for sets of relations.

COROLLARY 4.8 *Let $\mathcal{S}, \mathcal{T} \subseteq 2^{\mathcal{B}}$ be two sets such that path-consistency decides* CSPSAT$(\mathcal{T})$, *and let M be a refinement matrix of $\mathcal{S}$.* GET-REFINEMENTS $(\mathcal{S}, M)$ *returns M'. If for every $S \in \mathcal{S}$ there is a $T_S \in \mathcal{T}$ with $M'[S][T_S] = true$, then path-consistency decides* CSPSAT$(\mathcal{S})$.

If a suitable refinement matrix can be found, CHECK-REFINEMENTS can be used to immediately verify that reasoning over the given set of relations is tractable. One problem with this method is that the algorithms, though polynomial, are not very efficient. Especially for large sets of relations the algorithms are very slow. Fortunately, the algorithms are used for determining tractability of reasoning over sets of relations and not for the reasoning process itself. Renz, 2002 proposed a faster version of the algorithm which uses a refinement array instead of a refinement matrix which reduces the runtime of the algorithm to $O(n^4 \log n)$.

In the following subsection we show how the refinement method can be applied to RCC-8 for proving certain subsets to be tractable. For RCC-8 it will lead to a complete analysis of tractability by identifying all three maximal tractable subsets.

5.3 Applying the refinement method

The refinement method requires for any input set of relations $\mathcal{S} \subseteq 2^{\mathcal{B}}$ a subset $\mathcal{T} \subseteq 2^{\mathcal{B}}$ for which path-consistency is known to decide consistency and a refinement strategy $\mathcal{S} \Rightarrow \mathcal{T}$. Assuming that a set $\mathcal{T}$ is known, the main tasks are to find a candidate set $\mathcal{S}$ and a refinement strategy, i.e., we have to find for every relation of $\mathcal{S}$ a relation of $\mathcal{T}$ and apply the refinement algorithm using the different refinement strategies.

Candidate sets $\mathcal{S}_i$ can be found by using the closure method and the known NP-hard relations. Renz, 1999 identified candidate sets for RCC-8 by computing the largest subsets of RCC-8 that contain the basic relations, the universal relation, are closed under the operators and do not contain any of the known NP-hard relations, i.e., the relations that can be used for the NP-hardness proofs. This resulted in only three candidate sets (which are called $\mathcal{C}_8$, $\mathcal{Q}_8$ and the

already known maximal tractable subset $\widehat{\mathcal{H}}_8$) which can be tested using the refinement method, provided that a refinement strategy can be found.

One way of finding a refinement strategy is to use a greedy method of extending partial refinement strategies by first refining only one or a few relations, fill the refinement matrix/array using the refinement algorithm and if no inconsistency occurs add some more refinements. This can be repeated until a working refinement strategy can be found or until it is shown that no refinement strategy exists. It might also be possible that a particular refinement strategy, the *identity refinement strategy*, is applicable. The identity refinement strategy refines each relation $R \in \mathcal{S}$ to the relation $R' = R \setminus ID$, where $ID \in \mathcal{B}$ is the identity relation. Renz, 1999 observed that all relations that are contained in the two candidate sets for RCC-8 which are not contained in $\widehat{\mathcal{H}}_8$ can be refined to relations of $\widehat{\mathcal{H}}_8$ by removing the identity relation. It turned out that applying the refinement algorithm to the candidate sets of RCC-8 leads to refinement matrices that contain a basic relation for each relation of the candidate sets. This shows that each of these candidate sets can be refined to the set of basic relations and, therefore, that the candidate sets are tractable and can be decided by the path-consistency algorithm. Renz, 1999 also applied the identity refinement matrix to the known maximal tractable subset ORD-Horn of the interval algebra (Nebel and Bürckert, 1995) and it turns out that the refinement method also works for the interval algebra. The refinement method has recently been extended to identify tractable subsets automatically given only the composition table of the relations (Renz, 2007).

Now we have all the tools for identifying (maximal) tractable subsets of a system of spatial relations. In the next subsection we show how these sets can be used for finding fast solutions to intractable CSPSAT instances.

6. Practical efficiency of reasoning methods

In the previous section we described how to find tractable subsets of the usually NP-hard spatial calculi. For most of the tractable subsets path-consistency or even simpler methods are sufficient for deciding consistency, so except for very large instances or for calculi over a large set of relations, there are usually no efficiency problems when considering instances that contain only relations of a tractable subset. Efficiency problems occur, however, if we go outside the tractable subsets and enter the NP-hard territory. As we will see later, instances of an NP-hard problem can often be solved very fast in practice. There are basically four reasons for this. The first one is that the interleaved applications of path-consistency during the backtracking search is often very powerful and already eliminates many labels that cannot lead to a solution. The second reason is that large tractable subsets reduce the size of the backtracking search tree by several orders of magnitude. This results from the possibility of splitting relations into tractable sub-relations instead of splitting them into all contained basic relations. The third reason is that different heuristics and strategies can be

applied for solving hard instances. Often it is the case that there is a heuristic for which a hard instances turns out to be easy. So the more heuristics and strategies are available the higher is the likelihood that one of them can solve an instance fast. The last reason is the observation that most instances outside the phase-transition region are in almost all cases very easy to solve. In the following we will discuss these points in more detail and show results from an empirical investigation of the practical efficiency of RCC-8.

6.1 Generating Test Instances

In order to test the practical efficiency of reasoning algorithms, it is necessary to generate a large number of test instances. Ideally these should be real instances of existing applications. If such an application is not available, instances have to be generated systematically or randomly. Since many instances are easy to solve, it is important to try to generate instances that are as hard as possible. When randomly generating instances, there is usually a parameter that produces a phase-transition of the probability of satisfiability of the generated instances, i.e., when increasing the value of the parameter, the probability changes from almost 1 to almost 0 (or vice versa) within a very small range of the parameter (see Fig. 4.8). Almost all instances outside the phase-transition region are very easy to solve while the phase-transition region contains most hard instances (Cheeseman et al., 1991). The most useful instances for empirical study of reasoning algorithms can therefore be found in and around the phase-transition region, which has to be empirically determined.

For randomly generated RCC-8 instances it turned out that one phase-transition is induced by the degree d of nodes, i.e., how many edges for each node of the constraint graph are randomly instantiated on average (Renz and Nebel, 2001). The phase-transition turns out to be around $d = 10$. Another way of generating hard instances is to randomly generate instances that contain only relations that are outside the tractable subsets. This, however, is mainly for testing the behaviour of the algorithms in extreme cases and is not very representative for practical purposes for which it might better to analyse a uniform distribution of the relations. Another important factor when generating random instances is to make sure that the instances are not trivially flawed (Achlioptas et al., 1997), i.e., the probability that small inconsistent sub-CSPs, such as inconsistent triples, are contained in the instances should not be high and should not determine the phase-transition.

The following empirical results are taken from Renz and Nebel, 2001 and show how randomly generated RCC-8 instances can be solved very efficiently. The random RCC-8 instances were generated according to the model $A(n, d, l)$, where n is the number of variables, d the average degree and l the average number of base relations per relation. The relations were selected among all

RCC-8 relations, for RCC-8 $l = 4.0$ means that all relations are selected with equal probability.

6.2 Testing Algorithms

When testing algorithms on the generated instances, several properties are interesting and should be observed. One is of course the time it takes to solve the instances, but it is also important to compare the number of nodes of the backtracking search space that were visited while solving instances. This value is important for comparing algorithms on different machines as the run-time differs from machine to machine and also depends on other factors such as the load and the available memory of the machine used for the test. Instead of using only the average values (runtime, visited nodes, etc.) we also look at different percentiles, i.e., we order the values and look at the values of the elements at position 50%, 70%, 90%, or 99%. Since we are dealing with an NP-complete problem for which some instances take a very long time to solve, taking the average only would be too erratic. In the following we mainly look at 99% percentile instances as these give a good indication of the performance for the hardest among the instances.

6.3 Effect of using large tractable subsets

A very important factor in obtaining more efficient solutions to instances of an NP-hard spatial reasoning problem is the use of large tractable subsets of the NP-hard set of relations. The backtracking algorithms split each constraint into sub-constraints that contain only relations of a tractable subset where each split spans a new subtree of the search space. Using large tractable subsets makes it possible to split the constraints into fewer sub-constraints, thus reducing the number of subtrees and the size of the search space. This can be measured in terms of the average branching factor of a search tree. For RCC-8, using the set of basic relations for splitting the constraints leads to an average branching factor of $b = 4$ which corresponds in this case to the average number of basic relations in each of the 256 RCC-8 relations. For the maximal tractable subsets $\widehat{\mathcal{H}}_8$, $\mathcal{Q}_8$, and $\mathcal{C}_8$, the average branching factors are $b = 1.4375$, $b = 1.516$, and $b = 1.523$, respectively. The average size of the search spaces can be computed as $b^{(n^2-n)/2}$. As can be seen in Table 4.3 this results in considerably smaller search spaces. This however is not fully reflected in the empirical results because of the effect of the interleaved applications of the path-consistency algorithm at each node of the search tree which eliminates inconsistent relations from the constraints and has a similar effect of reducing the search space. Both methods together, path-consistency and large tractable subsets, already lead to quite impressive results for solving randomly generated RCC-8 instances. In Fig. 4.11 we see the 99% percentile running times for solving instances of the

#regions	$\mathcal{B}(4.0)$	$\widehat{\mathcal{B}}(2.5)$	$\widehat{\mathcal{H}}_8$ (1.4375)
5	10^6	9537	37
7	4.4×10^{12}	2.3×10^8	2040
10	1.2×10^{27}	8.1×10^{17}	10^7
20	2.5×10^{114}	4.1×10^{75}	8.8×10^{29}

Table 4.3. Average size of the search space depending on the number of variables and the branching factor of the split set.

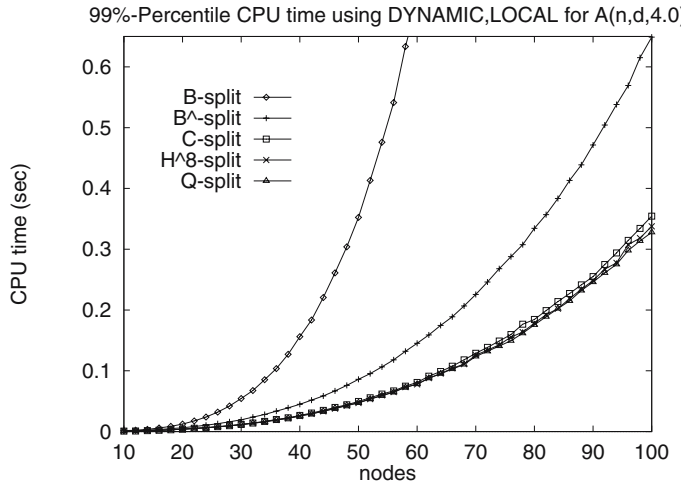


Figure 4.11. 99% percentile running times for solving RCC-8 instances of the phase-transition region using different tractable subsets ($d = 8.0$ to $d = 10.0$, 2,500 instances per data point).

phase-transition region using different tractable subsets. The maximal tractable subset lead to considerably faster solutions but not as much faster as suggested by table 4.3

6.4 Effect of different heuristics

Another factor for obtaining faster solutions is to use different heuristics for choosing the path through the search space. There are two positions in the backtracking algorithms where a heuristic choice can be made. One is the order in which the constraints are selected, the other choice is the order of the sub-relations when splitting a constraint. For both choices we can apply different heuristics which influence the search space and the path through the search space. It is clear that the choice of the heuristics has more effect on consistent instances. In order to determine that an instance is consistent, it is sufficient to find one path from the root of the search tree to a consistent leaf.

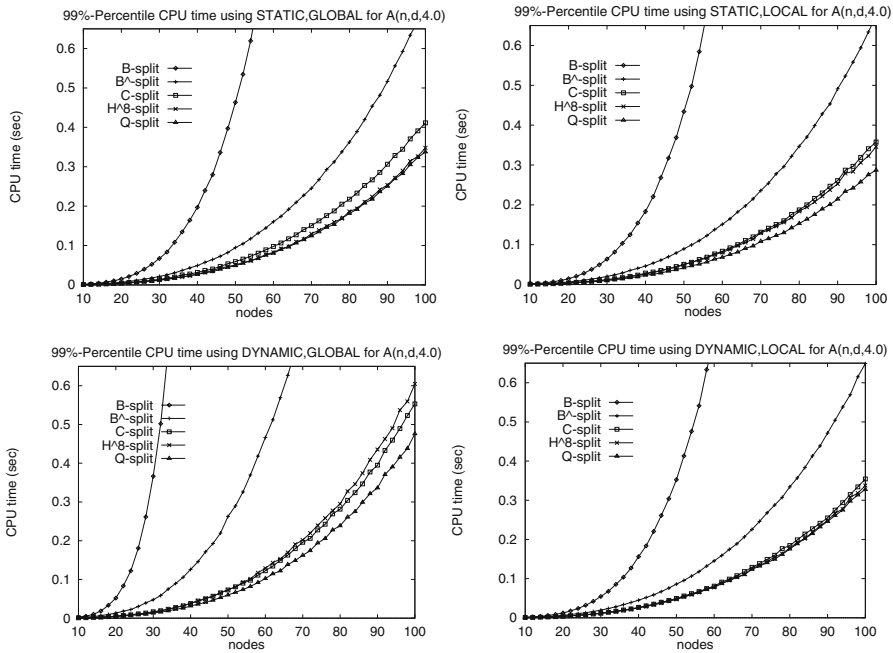


Figure 4.12. Percentile 99% CPU time of the different heuristics for solving $A(n, d, 4.0)$ ($d = 8.0$ to $d = 10.0$, 2,500 instances per data point).

So if the perfect heuristic choice is made at all nodes, any consistent instance can be solved without backtracking. For inconsistent instances, all possible leafs of the search tree must be inconsistent, so the fastest way to determine inconsistency is when this can be detected early on in the search tree. We chose two different heuristics for the ordering of constraints and two for the ordering of sub-relations (Nebel, 1997)

static/dynamic: Constraints are processed according to a heuristic evaluation of their constrainedness which is determined *statically* before the backtracking starts or *dynamically* during the search.

local/global: The evaluation of the constrainedness is based on a *local* heuristic weight criterion or on a *global* heuristic criterion (van Beek and Manchak, 1996).

In Fig. 4.12 the 99% percentiles are shown for the different combinations of heuristics, the second column of Table 4.4 shows the number of hard instances for each combinations. Hard instances are considered to be those that cannot be solved by using 10.000 visited nodes in the search space. It can be seen that although some combinations are better than others, they are all quite successful

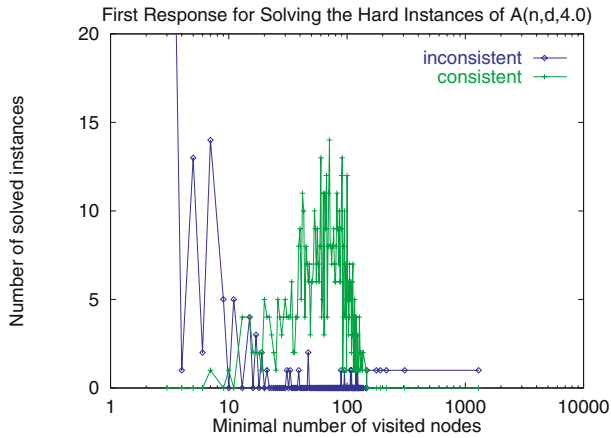


Figure 4.13. Fastest solution of the hard instances when running all heuristics in parallel.

and the differences are not enormous. Their real advantage is described in the following section.

6.5 Effects of combining different strategies

We denote as a strategy a choice of tractable subset for splitting, a heuristic for constraint selection and a heuristic for sub-relation ordering. As described in the previous section, every consistent instance can be solved without backtracking if the right heuristic choice is made at each node. Therefore it is not surprising that some instances can be solved faster by one strategy while other instances are solved faster by other strategies. This means that it might be possible to solve more instances efficiently by combining different strategies than by each strategy alone. We tested this hypothesis by running all strategies on the set of all hard instances identified in the experiment described above. It turns out that almost all of these hard instances can be solved by at least one strategy (see Table 4.4). We also looked at which strategy gives the first response, i.e., which strategy solves each instance fastest, which is shown in the same table. In most cases, the first response comes very fast, usually with less than 300 visited nodes in the search space (see Fig. 4.13). It is surprising that the inconsistent instances can be solved particularly fast which shows a clear advantage of the method of combining different strategies to random methods with restarts. Random methods are actually completely useless for inconsistent instances because these methods are not complete. In order to push our methods even further, we also looked at how well the different strategies complement each other, and tried to find the combination of strategies which solves the instances with the least accumulated number of nodes. It turns

Heuristics	# Hard Instances	$A(n, d, 4.0)$ Solved Instances	1. Response
$\widehat{\mathcal{H}}_8$/sta/loc	64	91.88%	19.80%
$\widehat{\mathcal{H}}_8$/sta/glo	42	94.67%	12.56%
$\widehat{\mathcal{H}}_8$/dyn/loc	52	93.40%	24.37%
$\widehat{\mathcal{H}}_8$/dyn/glo	100	87.31%	13.58%
$\mathcal{C}_8$/sta/loc	81	89.72%	6.35%
$\mathcal{C}_8$/sta/glo	58	92.64%	5.20%
$\mathcal{C}_8$/dyn/loc	78	90.10%	5.96%
$\mathcal{C}_8$/dyn/glo	108	86.63%	6.60%
$\mathcal{Q}_8$/sta/loc	81	89.72%	9.77%
$\mathcal{Q}_8$/sta/glo	54	93.15%	12.06%
$\mathcal{Q}_8$/dyn/loc	74	90.61%	10.15%
$\mathcal{Q}_8$/dyn/glo	104	86.80%	12.82%
$\widehat{\mathcal{B}}$/sta/loc	68	91.37%	1.40%
$\widehat{\mathcal{B}}$/sta/glo	89	88.71%	1.27%
$\widehat{\mathcal{B}}$/dyn/loc	70	91.12%	0.89%
$\widehat{\mathcal{B}}$/dyn/glo	162	79.44%	0.89%
$\mathcal{B}$/sta/loc	163	79.31%	0.51%
$\mathcal{B}$/sta/glo	222	71.83%	0.25%
$\mathcal{B}$/dyn/loc	209	73.48%	0.51%
$\mathcal{B}$/dyn/glo	(303)	–	0.13%
combined	788	99.87%	

Table 4.4. The second column shows the number of hard instances for each heuristic, there are 788 hard instances in total. Column three shows the percentage of solved hard instances for each heuristic and column four the percentage of first response when orthogonally running all heuristics. Note that sometimes different heuristics are equally fast. Therefore the sum is more than 100%.

out that four strategies ($\widehat{\mathcal{H}}_8$/*static*/*global*, $\widehat{\mathcal{H}}_8$/*dynamic*/*local*, $\mathcal{C}_8$/*dynamic*/*local*, $\widehat{\mathcal{B}}$/*static*/*local*) complement each other particularly well. By combining these four strategies, almost all instances in the phase transition region can be solved by restricting the combined number of visited nodes to a value which is linear in the size of the instances. We tested this for CSPs up to a size of 500 variables, i.e., CSPs with about 25.000 relations. At that point the increased run-time of the interleaved path-consistency computations turned out to be the limiting factor.

6.6 Discussion

We have seen that even though spatial reasoning with RCC-8 is an NP-complete problem, we were able to solve almost all of the hardest instances

identified in our experiments in reasonable time. This is only possible through the use of the maximal tractable subsets that we identified by a theoretical analysis of the reasoning problem. For RCC-8 this turns out to work particularly well, which is due to the existence of three different maximal tractable subsets for RCC-8 but also because RCC-8 is a rather small algebra. Empirical studies for Allen's interval algebra (Allen, 1983) which has 13 basic relations but only one maximal tractable subset which contains all basic relations, show that reasoning is still much more efficient in practice when using the maximal tractable subclass than without (Nebel, 1997), but the overall practical efficiency was not as impressive as for RCC-8. Nevertheless, identifying maximal tractable subsets that contain all basic relations is an essential part if more efficient solutions to an NP-complete spatial or temporal reasoning problem are to be found. Although attempts have been made to identify maximal tractable subsets that do not contain all basic relations (Krokhin et al., 2003), these subsets cannot be used for obtaining more efficient solutions to the general NP-complete problem as it is not possible to split each relation into members of these maximal tractable subset. Another important finding is that combining different strategies leads to much better results than trying to optimise one strategy. In that respect we can conclude that the more strategies the better. This includes analysing different heuristics as well as using different tractable subsets which even includes using subsets of tractable subsets.

7. Combination of spatial calculi

There has been a large amount of research on qualitative spatial calculi, a fraction of it has been described in this chapter, and more in other chapters of this book. The usefulness of these research efforts, however, largely depends on how well this research can make its way into practical applications. Without a doubt, space is one of the fundamental aspects of our daily life and of our physical world, and therefore qualitative spatial representation and reasoning should be an essential part in many applications. It is remarkable, however, that up to now there are relatively few real applications. One reason for this lack of applications is that research has mainly focused on understanding and analysing single, isolated aspects of space, like distance, direction, or topology. The spatial calculi we presented so far all fall into this "single aspect" category.

On the other hand, almost all possible applications require different aspects of space and not only topology or only direction. Future research on qualitative spatial representation and reasoning should focus strongly on combining different aspects of space, on developing and analysing spatial calculi over different aspects of space, and on methods for dealing with these calculi. In this chapter we will present some promising first attempts in this direction.

The first approach is by Gerevini and Renz, 2002, who combine topology and size information and who introduce several modifications of the existing constraint algorithms for dealing with different kinds of constraints. A second approach is by Renz, 2001, who combines directional and topological information for one-dimensional intervals by adding direction of intervals to the interval algebra.

7.1 Different ways of combining multiple aspects of space

Constraint-based approaches in principle support the use of different kinds of constraints if they work on the same domains. This is relatively straightforward if finite domains are used where the constraint algorithms manipulate the domains of the variables. For qualitative spatial reasoning where infinite domains are used and constraint algorithms work on relation-symbols instead of restricting domains, this turns out to be a difficult problem. The reason for this is that relations over one aspect are not independent of relations of another aspect. For example if the distance between two objects is large, they cannot overlap. If one object is contained in the other one it must be smaller. These are two simple examples which show that topology is neither independent from direction nor from size. These restrictions and dependencies must be enforced on the relational level and must therefore be analysed when developing a combined calculus and must be precomputed like a composition table.

One way of developing a calculus for multiple aspects of space is to take the relations for each aspect, for example two sets of basic relations $\mathcal{R} = \{R_1, \ldots, R_n\}$ and $\mathcal{S} = \{S_1, \ldots, S_m\}$, and form new relations as the cross product $\mathcal{R} \times \mathcal{S}$. Some of the new relations will be empty and can be removed. The advantage of this approach is that the dependencies of the different aspects are implicitly encoded in the new composition table and that all the existing reasoning algorithms can be used. The disadvantage is the large number of relations that result from this approach (which makes reasoning and also analysing the combined calculus very time and space consuming). An example for this approach is by Pujari et al., 1999 in the area of temporal reasoning where the interval algebra is combined with relative durations of intervals. Another example is by Renz, 2001, who added direction of intervals to the interval algebra.

An alternative approach is to treat the different aspects separately and to develop new reasoning algorithms for combining different sets of constraints and their dependencies. Different aspects and different granularities can then be added in a modular way without having an explosion in the number of relations. One problem here is how to keep track of the interactions between the different sets of relations and their interactions, i.e, how can relations of different sets be composed, intersected etc. This approach will be further discussed in the

r		$Sizerel(r)$	r		$Sizerel(r)$	s		$Toprel(s)$
TPP	$\models$	$<$	DC	$\models$	?	$=$	$\models$	DC, EC, PO, EQ
NTPP	$\models$	$<$	EC	$\models$	?	$>$	$\models$	DC, EC, PO, TPP^{-1}, NTPP^{-1}
TPP^{-1}	$\models$	$>$	PO	$\models$	?	$<$	$\models$	DC, EC, PO, TPP, NTPP
NTPP^{-1}	$\models$	$>$	EQ	$\models$	$=$			

Table 4.5. Interdependencies of basic RCC-8 relations (r) and basic $\mathcal{QS}$ relations (s).

following section where we take the combination of topology and qualitative size as an example.

In any case, it is essential that different aspects can only be combined when they use the same underlying spatial entities, such as points or regular regions.

7.2 Combining topological and size information

When having two sets of basic relations $\mathcal{A}$ and $\mathcal{B}$ over a domain $\mathcal{D}$ where both of them split $\mathcal{D} \times \mathcal{D}$ exhaustively, it is clear that the relations of the two sets taken together are not pairwise disjoint and, hence, cannot be independent of each other. Instead of looking at the intersections of all the relations and to treat them as a new set of JEPD relations, we will present in this section methods for taking the sets separately and propagating their interactions. For this it is necessary to first look at all the interactions that can possibly occur. As an example we take the work by Gerevini and Renz, 2002, who combined RCC-8 with qualitative size relations. Given a set V of spatial region variables, a set of $\mathcal{QS}$-constraints over V is a set of constraints of the form $size(x)\ S\ size(y)$, where $S \in \mathcal{QS}$, $size(x)$ is the size of the region x, $size(y)$ is the size of the region y, and $x, y \in V$. $\mathcal{QS} = \{<, >, =, \leq, \geq, \neq, <=>\}$. Their interactions are rather simple and are mainly due to the fact that regions which are contained in other regions must be smaller than the containing region. All interactions can be found in Table 4.5. $Sizerel(r)$ is the qualitative size relation entailed by an RCC-8 relation r, while $Toprel(s)$ is the RCC-8 relation entailed by a qualitative size relation.

Now we consider pairs of relations as new relations, i.e., we consider constraints of the form xRy where $R = \langle R_a, R_b \rangle$ and $R_a \in \mathcal{A}$ and $R_b \in \mathcal{B}$. This is equivalent to having two sets of constraints Θ and Σ over the same set of variables and the same domain where Θ contains the RCC-8 constraints and Σ the qualitative size constraints. If both sets are independently consistent, it is clear that the two sets taken together are not necessarily consistent too as their interactions must be considered.

A natural method for deciding the consistency of a set of RCC-8 constraints and a set of $\mathcal{QS}$-constraints, would be to first extend each set of constraints with the constraints entailed by the other set, and then independently check

the consistency of the extended sets by using a path-consistency algorithm. However, as the example below shows, this method is not complete for $\widehat{\mathcal{H}}_8$ constraints.

Another possibility would be to compute the strongest entailed relations (minimal relations) between each pair of variables before propagating constraints from one set to the other. However, this method has the disadvantage that it is computationally expensive, as the best known algorithm for computing the minimal network of a set of constraints over either $\widehat{\mathcal{H}}_8$, $\mathcal{C}_8$ or $\mathcal{Q}_8$ requires $O(n^5)$ time.

Finally, a third method could be based on iteratively using path-consistency as a preprocessing technique and then propagating the information from one set to the other. A similar method is used by Ladkin and Kautz to combine qualitative and metric constraints in the context of temporal reasoning (Kautz and Ladkin, 1991). Note that imposing path-consistency is sufficient for consistency checking of a set of constraints over $\widehat{\mathcal{H}}_8$, $\mathcal{C}_8$, $\mathcal{Q}_8$, and $\mathcal{QS}$, but is incomplete for computing the minimal relations (van Beek, 1992; Renz and Nebel, 1999). The following example shows that the information would need to be propagated more than once, and furthermore it is not clear whether in general this method would be complete for detecting inconsistency.

EXAMPLE 4.9 *Consider the set Θ formed by the following $\widehat{\mathcal{H}}_8$ constraints*

$$x_0\{\mathsf{TPP}, \mathsf{EQ}\}x_2,\ x_1\{\mathsf{TPP}, \mathsf{EQ}, \mathsf{PO}\}x_0,\ x_1\{\mathsf{TPP}, \mathsf{EQ}\}x_2,\ x_4\{\mathsf{TPP}, \mathsf{EQ}\}x_3,$$

and the set Σ formed by the of following $\mathcal{QS}$-constraints

$$size(x_0) < size(x_2),\ size(x_3) \leq size(x_1),\ size(x_2) \leq size(x_4).$$

We have that Θ and Σ are independently consistent, but their union is not consistent. Moreover, the following propagation scheme does not detect the inconsistency: (a) enforce path-consistency to Σ and Θ independently; (b) extend Σ with the size constraints entailed by the constraints in Θ; (c) extend Θ with the topological constraints entailed by the constraints in Σ; (d) enforce path-consistency to Θ and Σ again. In order to detect that $\Theta \cup \Sigma$ is inconsistent, we need an additional propagation of constraints from the topological set to the size set.

Instead of directly analysing the complexity and completeness of the propagation scheme illustrated in the previous example, Gerevini and Renz, 2002 proposed a new method for dealing with combined topological and qualitative size constraints. In particular, they propose an $O(n^3)$ time and $O(n^2)$ space algorithm, BIPATH-CONSISTENCY, for imposing path-consistency to a set of constraints in $\mathsf{RCC\text{-}8} \cup \mathcal{QS}$. BIPATH-CONSISTENCY solves CSPSAT for any input set Θ of topological constraints in either $\widehat{\mathcal{H}}_8$, $\mathcal{C}_8$ or $\mathcal{Q}_8$, combined with

Algorithm: BIPATH-CONSISTENCY
Input: A set Θ of RCC-8 constraints, and a set Σ of $\mathcal{QS}$-constraints over the variables $x_1, x_2, \ldots, x_n$ of Θ.
Output: `fail`, if $\Sigma \cup \Theta$ is not consistent; path-consistent sets equivalent to Σ and Θ, otherwise.

1. Q $\leftarrow \{(i,j) \mid i < j\}$; ($i/j$ indicates the i-th/j-th variable of Θ.)
2. *while* Q $\neq \emptyset$ *do*
3. select and delete an arc (i,j) from Q;
4. *for* $k \neq i, k \neq j$ ($k \in \{1...n\}$) *do*
5. *if* BIREVISION(i,j,k) *then*
6. *if* $R_{ik} = \emptyset$ *then* return `fail`
7. *else* add (i,k) to Q;
8. *if* BIREVISION(k,i,j) *then*
9. *if* $R_{kj} = \emptyset$ *then* return `fail`
10. *else* add (k,j) to Q.

Function: BIREVISION(i,k,j)
Input: three region variables i, k and j
Output: true, if R_{ij} is revised; false otherwise.
Side effects: R_{ij} and R_{ji} revised using the operations $\cap$ and $\circ$ over the constraints involving i, k, and j.

1. *if* one of the following cases hold, *then* return false:
   (a) $Toprel(s_{ik}) \cap t_{ik} = U_t$ *and* $Sizerel(t_{ik}) \cap s_{ik} = U_s$,
   (b) $Toprel(s_{kj}) \cap t_{kj} = U_t$ *and* $Sizerel(t_{kj}) \cap s_{kj} = U_s$
2. oldt := t_{ij}; olds := s_{ij};
3. $t_{ij} := (t_{ij} \cap Toprel(s_{ij})) \cap ((t_{ik} \cap Toprel(s_{ik})) \circ (t_{kj} \cap Toprel(s_{kj})))$;
4. $s_{ij} := (s_{ij} \cap Sizerel(t_{ij})) \cap ((s_{ik} \cap Sizerel(t_{ik})) \circ (s_{kj} \cap Sizerel(t_{kj})))$;
5. *if* $s_{ij} \neq$ olds *then* $t_{ij} := (t_{ij} \cap Toprel(s_{ij}))$;
6. *if* (oldt = t_{ij}) *and* (olds = s_{ij}) *then* return *false*;
7. $t_{ji} := Converse(t_{ij})$; $s_{ji} := Converse(s_{ij})$;
8. return *true*.

Figure 4.14. BIPATH-CONSISTENCY.

any set of size constraints in $\mathcal{QS}$ involving the variables of Θ. Thus, despite this framework is more expressive than a purely topological one over the same set of relations (and therefore has a larger potential applicability), the problem of deciding consistency can be solved without additional worst-case cost.

BIPATH-CONSISTENCY is a modification of Vilain and Kautz' path-consistency algorithm (Vilain and Kautz, 1986; Vilain et al., 1989) as described by Bessière, 1996, which in turn is a slight modification of Allen's algorithm (Allen, 1983). The main novelty of the algorithm is that BIPATH-CONSISTENCY operates on a graph of *pairs* of constraints. The vertices of the graph are constraint variables, which in our context correspond to spatial regions. Each edge of the graph is labelled by a pair of relations formed by a topological relation in RCC-8 and a size relation in $\mathcal{QS}$. The function BIREVISION(i, k, j) has the same role as the function REVISE used in path consistency algorithms for constraint networks (e.g., Mackworth, 1977). The main difference is that BIREVISION(i, k, j) considers pairs of (possibly interdependent) constraints, instead of single constraints.

A formal description of BIPATH-CONSISTENCY is given in Fig. 4.14, where R_{ij} is a pair formed by a relation t_{ij} in RCC-8 and a relation s_{ij} in $\mathcal{QS}$; $R_{ij} = \emptyset$ when $t_{ij} = \emptyset$ or $s_{ij} = \emptyset$; U_t indicates the universal relation in RCC-8 and U_s the universal relation in $\mathcal{QS}$.

Gerevini and Renz, 2002 prove soundness and completeness of BIPATH-CONSISTENCY for the maximal tractable subsets of RCC-8 combined with qualitative size relations.

THEOREM 4.10 *Given a set Θ of constraints in either $\widehat{\mathcal{H}}_8$, $\mathcal{C}_8$ or $\mathcal{Q}_8$, and a set Σ of constraints in $\mathcal{QS}$ involving variables in Θ, consistency of $\Theta \cup \Sigma$ can be decided using the* BIPATH-CONSISTENCY *algorithm in $O(n^3)$ time and $O(n^2)$ space, where n is the number of variables involved in Θ and Σ.*

Using the BIPATH-CONSISTENCY algorithm, combined sets of constraints can be solved in cubic time just like the normal path-consistency algorithm for each of the two sets alone, i.e., they can be solved without additional worst-case cost. Soundness and completeness of BIPATH-CONSISTENCY do not hold automatically and has to be proved for each combination of different relations anew. Sometimes, however, the computational properties of combined calculi can be more favourable than both of them alone if the interactions with the other type of relations refines relations that make deciding consistency NP-hard to relations for which it is tractable. As can be seen in Table 4.5, whenever we have a definite qualitative size constraint $size(x)Ssize(y)$ with $S \in \{<, >, =\}$ and this constraint is combined with an RCC-8 constraint xRy resulting in $xR'y$, then R' will contain basic relations of at most one of the sets $\{\mathsf{TPP}, \mathsf{NTPP}\}$, $\{\mathsf{TPP}^{-1}, \mathsf{NTPP}^{-1}\}$, $\{\mathsf{EQ}\}$ and possibly some relations of the set $\{\mathsf{DC}, \mathsf{EC}, \mathsf{PO}\}$. In other words, any relation of $R \in$ RCC-8 $\setminus \widehat{\mathcal{H}}_8$ will be refined to a relation $R' \in \widehat{\mathcal{H}}_8$ and therefore it is possible that sets of constraints for which it is NP-hard to decide consistency become tractable after adding constraints over a different set of relations.

The opposite is of course also possible, that combining two calculi results in a calculus that has a higher complexity than both alone. This is the case for another

combination that Gerevini and Renz analysed, namely, combining RCC-8 with metric size information. They considered metric size constraints of different kinds, metric relative size constraints $size(x)R\alpha \cdot size(y)$ where α is a positive rational number, size difference constraints $size(x) - size(y) \in I$ where I is a continuous interval of rational numbers, or domain size constraints $size(x) \in I$. The main difference of combining RCC-8 with metric size information as compared to qualitative size information is that it is possible to express that a set of regions completely fills another region. Combining RCC-8 with any of these metric size calculi, which are all independently tractable, leads to NP-hardness even when combined with only the RCC-8 basic relations. Without the PO relation, i.e., considering only the 7 other RCC-8 basic relations, the combination is tractable though.

This was an example of how different calculi can be combined. Future research effort within qualitative spatial representation and reasoning should deal with modularising different aspects and different granularities in a similar way to how topology and size was combined here, studying their interactions and developing algorithms for reasoning about combined calculi. Then different applications could use the modules that are needed for the particular application, the interactions between the different modules and combine them using algorithms like BIPATH-CONSISTENCY. If possible, these combinations should have favourable computational properties and should enable efficient solutions.

7.3 Combining topological and directional information for intervals

In this section we give an example for a combination of two aspects of space that relies upon forming new relations out of two given sets of relations. One of the two aspects of space we are looking at is the interval algebra (IA) (Allen, 1983), which was originally defined for temporal reasoning. However, for applications that can require only a one-dimensional spatial representation, it makes sense to use the interval algebra. Some possible applications are from the area of traffic management. Roads and railway lines can be regarded as one-dimensional routes, but also air and sea traffic mainly operates on given routes. A single route can be represented as a one-dimensional space and the vehicles on a route as intervals. The main difference of vehicles on a route to the interval algebra is that vehicles and also routes have a direction. This can be captured by extending the interval algebra with direction of intervals (Renz, 2001). Direction in a one-dimensional space is quite simple as there are only two directions, front and back, or same direction and different direction.

A straightforward way for dealing with directed intervals would be to add additional constraints on the direction of intervals to constraints over the Interval

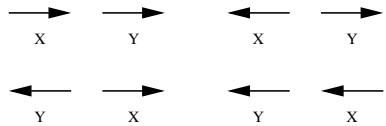


Figure 4.15. Four structurally different instantiations of the relation "x behind y" with directed intervals.

Algebra and treat the two types of constraints separately while propagating information from one type to the other (similar to what has been done in Gerevini and Renz, 1998.) We say that an interval has *positive direction* if it has the same direction as the underlying line and *negative direction* otherwise. So possible direction constraints could be unary constraints like "x has positive/negative direction" or binary constraints like "x and y have the same/opposite direction". This approach, however, is not possible since the Interval Algebra loses its property of being a relation algebra when permitting directed intervals. This can be easily seen when considering the "behind" relation of Fig. 4.15. The actual converse of "x behind y" is a subset of "y is behind or in front of x" which cannot be expressed within the Interval Algebra. If using "y is behind or in front of x" as the converse of "x behind y", whose converse is again "x is behind or in front of y", then applying the converse operation ($\cdot^{\smile}$) twice leads to a different relation than the original relation. This is a contradiction to one of the requirements of relation algebras ($R^{\smile\smile} = R$) (Ladkin and Maddux, 1994). This contradiction does not occur when we refine the "behind" relation into two disjoint sub-relations "behind$_=$" and "behind$_\neq$" where the subscript indicates that both intervals have the same ($=$) or opposite ($\neq$) direction. The converse of both relations is "in-front-of$_=$" and "behind$_\neq$", respectively. Applying the converse operation again leads to the original relations.

Since a relation algebra must be closed under composition, intersection, and converse, we have to make the same distinction also for all other IA relations. This leads to the definition of the directed intervals algerba DIA (Renz, 2001). It consists of the 26 basic relations given in Table 4.6, which result from refining each IA relation into two sub-relations specifying either same or opposite direction of the involved intervals, and of all possible unions of the basic relations. This gives a total number of 2^{26} DIA relations. Converse relations are given in the same table entry. If a converse relation is not explicitly given, the corresponding relation is its own converse. We denote the set of 26 DIA basic relations as $\mathcal{B}$. Then DIA $= 2^{\mathcal{B}}$. Complex relations which are the union of more than one basic relation $R_1, \ldots, R_k$ are written as $\{R_1, \ldots, R_k\}$. The union of all basic relations, the universal relation, is denoted $\{*\}$.

A DIA basic relation $R = I_d$ consist of two parts, the interval part I which is a spatial interpretation of the Interval Algebra and the direction part d which gives the mutual direction of both intervals, either $=$ or $\neq$. If a complex relation

Directed Intervals Basic Relation	Sym-bol	Pictorial Example
x behind$_=$ y	$\mathsf{b}_=$	-x->
y in-front-of$_=$ x	$\mathsf{f}_=$	-y->
x behind$_\neq$ y	$\mathsf{b}_\neq$	<-x- -y->
x in-front-of$_\neq$ y	$\mathsf{f}_\neq$	-x-> <-y-
x meets-from-behind$_=$ y	$\mathsf{mb}_=$	–x–>
y meets-in-the-front$_=$ x	$\mathsf{mf}_=$	–y–>
x meets-from-behind$_\neq$ y	$\mathsf{mb}_\neq$	<–x– –y–>
x meets-in-the-front$_\neq$ y	$\mathsf{mf}_\neq$	–x–> <–y–
x overlaps-from-behind$_=$ y	$\mathsf{ob}_=$	—x—>
y overlaps-in-the-front$_=$ x	$\mathsf{of}_=$	—y—>
x overlaps-from-behind$_\neq$ y	$\mathsf{ob}_\neq$	<—x— —y—>
x overlaps-in-the-front$_\neq$ y	$\mathsf{of}_\neq$	—x—> <—y—
x contained-in$_=$ y	$\mathsf{c}_=$	–x–>
y extends$_=$ x	$\mathsf{e}_=$	—y—>
x contained-in$_\neq$ y	$\mathsf{c}_\neq$	<–x–
y extends$_\neq$ x	$\mathsf{e}_\neq$	—y—>
x contained-in-the-back-of$_=$ y	$\mathsf{cb}_=$	–x–>
y extends-the-front-of$_=$ x	$\mathsf{ef}_=$	—y—>
x contained-in-the-back-of$_\neq$ y	$\mathsf{cb}_\neq$	<–x–
y extends-the-back-of$_\neq$ x	$\mathsf{eb}_\neq$	—y—>
x contained-in-the-front-of$_=$ y	$\mathsf{cf}_=$	–x–>
y extends-the-back-of$_=$ x	$\mathsf{eb}_=$	—y—>
x contained-in-the-front-of$_\neq$ y	$\mathsf{cf}_\neq$	<–x–
y extends-the-front-of$_\neq$ x	$\mathsf{ef}_\neq$	—y—>
x equals$_=$ y	$\mathsf{eq}_=$	—x—> —y—>
x equals$_\neq$ y	$\mathsf{eq}_\neq$	—x—> <—y—

Table 4.6. The 26 basic relations of the directed intervals algebra.

R consist of basic relations with the same direction part d, we can combine the interval parts and write $R = \{I^1, \ldots, I^k\}_d$ instead of $R = \{I^1_d, \ldots, I^k_d\}$. We write R_e (resp. R_n) in order to refer to the union of the interval parts of every sub-relation of a complex relation R where the direction part is $\{=\}$ (resp. $\{\neq\}$.) In this way, every DIA relation R can be written as $R = \{R_e\}_= \cup \{R_n\}_\neq$. DIA_I denotes the set of 2^{13} possible interval parts of DIA relations.

R	$\prec$	$\succ$	m	mi	o	oi	s	si	d	di	f	fi	$\equiv$
R^r	$\succ$	$\prec$	mi	m	oi	o	f	fi	d	di	s	si	$\equiv$
dia(R)	b	f	mb	mf	ob	of	cb	ef	c	e	cf	eb	eq

Table 4.7. IA basic relations R, their reverses R^r, and their spatial interpretations dia(R).

It is important to note that the spatial interpretation of the Interval Algebra was chosen in a way that the interval part of a relation xI_dy only depends on the direction of y and not on the direction of x. Therefore, if the direction of x is reversed, written as $\overline{x}$, then only the direction part changes, i.e., $xI_dy = \overline{x}I_{\neg d}y$. This would not be the case in a straightforward spatial interpretation of the original temporal relations. For instance, IA relations like "x started-by y" or "x finished-by y" depend on the direction of x. Instead, these relations are interpreted spatially as "x extends-the-front/back-of y" and "x contained-in-the-front/back-of y". This interpretation is independent of the direction of x. When all intervals have the same direction, both interpretations are equivalent. In order to transform the spatial and the temporal interval relations (independent of the direction of the intervals) into each other, we introduce two mutually inverse functions dia : IA $\mapsto$ DIA$_I$ and ia : DIA$_I$ $\mapsto$ IA, i.e., dia(ia$(R)) = R$ and ia(dia$(R)) = R$. The mapping is given in Table 4.7.

All relations of the directed intervals algebra are invariant with respect to the direction of the underlying line, i.e., when reversing the direction of the line, all relations remain the same. This is obviously not the case for the Interval Algebra, e.g., if x is before y and one reverses the direction of the time line, then x is after y. In order to transform DIA relations into the corresponding IA relations and *vice versa*, we introduce a unary *reverse* operator $(\cdot^r)$ on relations R such that R^r specifies the relation which results from R when reversing the direction of the underlying line. For all relations $R \in$ DIA we have that $R^r = R$. For IA relations, the reverse relation is given in Table 4.7. The reverse of a complex relation is the union of the reverses of the involved basic relations. The reverse of the composition ($\circ$) of two relations is equivalent to the composition of the reverses of the two involved relations, i.e., $(R \circ S)^r = R^r \circ S^r$. Applying the reverse operator twice results in the original relation, i.e., $R^{rr} = R$. Using the reverse operator we can also specify what happens with a relation xI_dy if only the direction of y is changed. Then the topological relation of the intervals stays the same, but the order changes, i.e., "front" becomes "behind"/"back" and *vice versa*. The mutual direction also changes. This can be expressed in the following way: $xI_dy = x\,\mathsf{dia}(\mathsf{ia}(I)^r)_{\neg d}\,\overline{y}$.

We now have all requirements for computing the composition ($\circ_d$) of DIA relations using composition of IA relations (denoted here by $\circ_{ia}$) as specified by Allen, 1983.

THEOREM 4.11 *Let R_p, S_q be* DIA *basic relations.*

1 *If $q = \{=\}$, then $R_p \circ_d S_q = \mathsf{dia}(\mathsf{ia}(R) \circ_{ia} \mathsf{ia}(S))_p$*

2 *If $q = \{\neq\}$, then $R_p \circ_d S_q = \mathsf{dia}(\mathsf{ia}(R)^r \circ_{ia} \mathsf{ia}(S))_{\neg p}$*

The composition of complex relations is as usual the union of the composition of the contained basic relations. It follows from the closedness of the Interval Algebra that DIA is closed under composition, intersection, converse, and reverse.

We have now obtained a new set of relations together with the necessary operations that cover the combination of the original two aspects. As opposed to the combination of topology and size, we can use the standard backtracking and path-consistency algorithms for reasoning about these relations. However, since we now have 26 basic relations, and 2^{26} possible subsets, it is much more difficult to analyse computational properties for these relations than it is for the interval algebra. Fortunately, it is partly possible to use the complexity results of the interval algebra in order to find complexity results for the directed interval algebra. NP-hardness of the directed intervals algebra follows immediately from NP-hardness of the interval algebra. But also tractable subsets can be identified by exploiting results for the interval algebra. Among them, most importantly, the set of DIA basic relations was shown to be tractable. In the proof of this result it is not important that all non-universal relations are basic relations, only that all non-universal relations consist of DIA basic relations with the same direction part. Therefore, tractability for the basic relations can be extend to the following result.

THEOREM 4.12 *Let $\mathcal{S}$ be a tractable subset of the Interval Algebra which is closed under the reverse operator. Then $\mathcal{S}^{\pm} = \{\mathsf{dia}(R)_{=} | R \in \mathcal{S}\} \cup \{\mathsf{dia}(R)_{\neq} | R \in \mathcal{S}\} \cup \{*\}$ is a tractable subset of the directed intervals algebra.*

Renz, 2001 showed that ORD-Horn, the only maximal tractable subset of the interval algebra that contains all basic relations is closed under the reverse operator. Therefore, the above theorem also applies to ORD-Horn. What's more, it was shown that path-consistency decides consistency for $\mathcal{H}^{\pm}$, the set of DIA relations which results from ORD-Horn (see Theorem 4.12).

THEOREM 4.13 *Path-consistency decides* CSPSAT*($\mathcal{H}^{\pm}$).*

Apart from these initial results the complexity of the directed intervals algebra is open and maximal tractable subsets have not yet been identified.

8. Conclusions

In this chapter we have shown how spatial information can be represented using constraint calculi. This is a natural way of using qualitative spatial calculi

and allows us to use methods and techniques developed for constraint satisfaction problems. We introduced qualitative spatial calculi, constraint satisfaction methods for reasoning over these calculi and presented methods for analysing the computational properties of spatial calculi. These methods are very general and can be used for different kinds of spatial calculi. We showed how a computational analysis can lead to very efficient reasoning even though the reasoning problems are NP-hard.

Using the presented methods we can specify a roadmap for how spatial calculi should be analysed in order to find efficient reasoning methods:

1 Define a useful set of basic relations $\mathcal{B}$ over a certain aspect of space, on a certain level of granularity and over a certain spatial domain.

2 Formally compute the composition table.

3 Try to find an NP-hardness proof for $\mathsf{CSPSAT}(2^{\mathcal{B}})$ using the method of polarity and clause constraints.

4 Try to show that $\mathsf{CSPSAT}(\mathcal{B})$ is tractable and that path-consistency decides consistency.

5 Identify larger tractable sets by applying the closure and the refinement methods. If possible identify maximal tractable subsets.

6 Using an empirical analysis, identify the combination of strategies that is most effective in solving instances of $\mathsf{CSPSAT}(2^{\mathcal{B}})$.

These methods work for calculi based on a single aspect of space. For most practical applications, however, it is necessary to combine more than one aspect of space. In this chapter we presented some example of how relations can be combined. One way is to form new relations that cover the combined aspects, another way is to treat the relations separately and to keep track of and propagate their interactions like it is done by Gerevini and Renz' bipath-consistency algorithm.

It is one of the main challenges of qualitative spatial reasoning to provide general methods for combining different calculi and for analysing the computational properties of combined calculi. What we presented here is only a small step in this direction and lots of future research is required. The goal of such an analysis could be a toolbox of calculi for different spatial and also temporal aspects on different granularities and efficient algorithms for their combinations. Each application could then pick the required sets of relations and the most efficient algorithms for combining these relations.

References

AAAI-96 (1996). *Proceedings of the 13th National Conference of the American Association for Artificial Intelligence*, Portland, OR. MIT Press.

Achlioptas, D., Kirousis, L., Kranakis, E., Krizanc, D., Molloy, M., and Stamatiou, Y. (1997). Random constraint satisfaction: a more accurate picture. In *3rd Conference on the Principles and Practice of Constraint Programming (CP'97)*, volume 1330 of *Lecture Notes in Computer Science*, pages 107–120. Springer Verlag.

Allen, James F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Asher, Nicholas and Vieu, Laure (1995). Towards a geometry of common sense: A semantics and a complete axiomatization of mereotopology. In IJCAI-95, 1995, pages 846–852.

Balbiani, Philippe, Condotta, Jean-François, and Farinas del Cerro, Luis (1998). A model for reasoning about bidimensional temporal relations. In Cohn et al., 1998, pages 124–130.

Balbiani, Philippe, Condotta, Jean-François, and Farinas del Cerro, Luis (1999a). A new tractable subclass of the rectangle algebra. In IJCAI-99, 1999, pages 442–447.

Balbiani, Philippe, Condotta, Jean-François, and Farinas del Cerro, Luis (1999b). A tractable subclass of the block algebra: constraint propagation and preconvex relations. In *Proccedings of the 9th Portuguese Conference on Artificial Intelligence*, pages 75–89.

Bennett, Brandon (1994). Spatial reasoning with propositional logic. In Doyle, J., Sandewall, E., and Torasso, P., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 4th International Conference*, pages 51–62, Bonn, Germany. Morgan Kaufmann.

Bessière, Christian (1996). A simple way to improve path-consistency in Interval Algebra networks. In AAAI-96, 1996, pages 375–380.

Biacino, Loredana and Gerla, Giangiacomo (1991). Connection structures. *Notre Dame Journal of Formal Logic*, 32(2):242–247.

Borgo, Stefano, Guarino, Nicola, and Masolo, Claudio (1996). A pointless theory of space based on strong connection and congruence. In Aiello, L.C., Doyle, J., and Shapiro, S.C., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 5th International Conference*, pages 220–229, Cambridge, MA. Morgan Kaufmann.

Cheeseman, Peter, Kanefsky, Bob, and Taylor, William M. (1991). Where the *really* hard problems are. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pages 331–337, Sydney, Australia. Morgan Kaufmann.

Clarke, Bowman L. (1981). A calculus of individuals based on connection. *Notre Dame Journal of Formal Logic*, 22(3):204–218.

Clarke, Bowman L. (1985). Individuals and points. *Notre Dame Journal of Formal Logic*, 26(1):61–75.

Clementini, Eliseo, di Felice, Paolino, and Hernandez, Daniel (1997). Qualitative representation of positional information. *Artificial Intelligence*, 95(2): 317–356.

Cohn, A.G., Schubert, L., and Shapiro, S.C., editors (1998). *Principles of Knowledge Representation and Reasoning: Proceedings of the 6th International Conference*, Trento, Italy.

Cohn, Anthony G., Bennett, Brandon, Gooday, John, and Gotts, Nicholas M. (1997). Representing and reasoning with qualitative spatial relations about regions. In Stock, O., editor, *Spatial and Temporal Reasoning*, pages 97–134. Kluwer, Dordrecht, Holland.

Cohn, Anthony G. and Varzi, Achille C. (1998). Connection relations in mereotopology. In *Proceedings of the 13th European Conference on Artificial Intelligence*, pages 150–154, Amsterdam, The Netherlands. Wiley.

Cohn, Anthony G. and Varzi, Achille C. (1999). Modes of connection. In *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, pages 299–314.

Cook, Stephen A. (1971). The complexity of theorem-proving procedures. In *Proc. 3rd Ann. ACM Symp. on Theory of Computing*, pages 151–158, New York. Association for Computing Machinery.

Cormen, Thomas H., Leiserson, Charles E., and Rivest, Ronald L. (1990). *Introduction to Algorithms*. MIT Press.

Dornheim, Christoph (1998). Undecidability of plane polygonal mereotopology. In Cohn et al., 1998.

Egenhofer, Max J. (1991). Reasoning about binary topological relations. In Günther, O. and Schek, H.-J., editors, *Proceedings of the Second Symposium on Large Spatial Databases, SSD'91*, volume 525 of *Lecture Notes in Computer Science*, pages 143–160. Springer-Verlag, Berlin, Heidelberg, New York.

Egenhofer, Max J., Clementini, Eliseo, and Felice, Paolino Di (1994). Topological relations between regions with holes. *International Journal of Geographical Information Systems*, 8(2):129–144.

Egenhofer, Max J. and Franzosa, Robert D. (1994). On the equivalence of topological relations. *International Journal of Geographical Information Systems*, 8(6):133–152.

Egenhofer, Max J. and Sharma, Jayant (1993). Assessing the consistency of complete and incomplete topological information. *Geographical Systems*, 1(1):47–68.

Forbus, Kenneth D., Nielsen, Paul, and Faltings, Boi (1987). Qualitative kinematics: A framework. In McDermott, J., editor, *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, Milan, Italy. Morgan Kaufmann.

Frank, Andrew U. (1991). Qualitative spatial reasoning about cardinal directions. In *Proceedings of the 7th Austrian Conference on Artificial Intelligence*, pages 157–167.

Freksa, Christian (1992). Using orientation information for qualitative spatial reasoning. In A.U. Frank, I. Campari, U. Formentini, editor, *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, volume 639 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, Heidelberg, New York.

Galton, Antony (1999). Mereotopology of discrete space. In Freksa, C. and Mark, D.M., editors, *Spatial information theory: Cognitive and computational foundations of geographic information science*, volume 1661 of *Lecture Notes in Computer Science*, pages 251–266, Berlin, Heidelberg, New York. Springer Verlag.

Garey, Michael R. and Johnson, David S. (1979). *Computers and Intractability—A Guide to the Theory of NP-Completeness*. Freeman, San Francisco, CA.

Gent, Ian P. and Walsh, Toby (1996). The satisfiability constraint gap. *Artificial Intelligence*, 81(1–2):59–80.

Gerevini, Alfonso and Renz, Jochen (1998). Combining topological and qualitative size constraints for spatial reasoning. In *Proceedings of the 4th International Conference on Principles and Practice of Constraint Programming*, Pisa, Italy.

Gerevini, Alfonso and Renz, Jochen (2002). Combining topological and size information for spatial reasoning. *Artificial Intelligence*, 137(1–2):1–42.

Golumbic, Martin C. and Shamir, Ron (1993). Complexity and algorithms for reasoning about time: A graph-theoretic approach. *Journal of the Association for Computing Machinery*, 40(5):1128–1133.

Gotts, Nicholas M. (1996). Using the RCC formalism to describe the topology of spherical regions. Technical Report 96-24, University of Leeds, School of Computer Studies.

Gotts, Nicholas M., Gooday, John M., and Cohn, Anthony G. (1996). A connection based approach to commonsense topological description and reasoning. *The Monist*, 79(1):51–75.

Goyal, Roop and Egenhofer, Max (2001). Similarity of direction relations. In Jensen, C., Schneider, M., Seeger, B. and Tsotras, V., editors, *Seventh international symposium on spatial and temporal databases*, volume 2121 of *Lecture Notes in Computer Science*, pages 36–55, Los Angeles, CA. Springer-Verlag.

Grigni, Michelangelo, Papadias, Dimitris, and Papadimitriou, Christos (1995). Topological inference. In IJCAI-95, 1995, pages 901–906.

Grzegorczyk, Andrzej (1951). Undecidability of some topological theories. *Fundamenta Mathematicae*, 38:137–152.

Guesgen, Hans (1989). Spatial reasoning based on Allen's temporal logic. Technical Report TR-89-049, ICSI, Berkeley, CA.

Hernàndez, Daniel (1994). *Qualitative Representation of Spatial Knowledge*, volume 804 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin, Heidelberg, New York.

Hernàndez, Daniel, Clementini, Eliseo, and di Felice, Paolino (1995). Qualitative distances. In *Spatial Information Theory: A Theoretical basis for GIS*, volume 988 of *Lecture Notes in Computer Science*, pages 45–58, Berlin, Heidelberg, New York. Springer-Verlag.

Hirsch, Robin (1999). A finite relation algebra with undecidable network satisfaction problem. *Bulletin of the IGPL.*

Hogg, Tad, Huberman, Bernardo A., and (eds), Colin P. Williams (1996). Special volume on frontiers in problem solving: Phase transistions and complexity. *Artificial Intelligence*, 81(1–2).

IJCAI-95 (1995). *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

IJCAI-99 (1999). *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, Sweden.

Isli, Amar and Moratz, Reinhard (1999). Qualitative spatial representation and reasoning: Algebraic models for relative position. Technical Report 284, Universität Hamburg, Fachbereich Informatik.

Johnson, David S. (1990). A catalog of complexity classes. In van Leeuwen, J., editor, *Handbook of Theoretical Computer Science, Vol. A*, pages 67–161. MIT Press.

Kautz, Henry A. and Ladkin, Peter B. (1991). Integrating metric and qualitative temporal reasoning. In *Proceedings of the 9th National Conference of the American Association for Artificial Intelligence*, pages 241–246, Anaheim, CA. MIT Press.

Krokhin, Andrei A., Jeavons, Peter, and Jonsson, Peter (2003). Reasoning about temporal relations: The tractable subalgebras of allen's interval algebra. *Journal of the ACM*, 50(5):591–640.

Ladkin, Peter B. and Maddux, Roger (1994). On binary constraint problems. *Journal of the Association for Computing Machinery*, 41(3):435–469.

Ladkin, Peter B. and Reinefeld, Alexander (1992). Effective solution of qualitative interval constraint problems. *Artificial Intelligence*, 57(1):105–124.

Ligozat, Gerard (1996). A new proof of tractability for ORD-Horn relations. In AAAI-96, 1996, pages 715–720.

Ligozat, Gerard (1998). Reasoning about cardinal directions. *Journal of Visual Languages and Computing*, 9:23–44.

Mackworth, Alan K. (1977). Consistency in networks of relations. *Artificial Intelligence*, 8:99–118.

Mackworth, Alan K. and Freuder, Eugene C. (1985). The complexity of some polynomial network consistency algorithms for constraint satisfaction problems. *Artificial Intelligence*, 25:65–73.

Montanari, Ugo (1974). Networks of constraints: fundamental properties and applications to picture processing. *Information Science*, 7:95–132.

Montello, Daniel R. (1993). Scale and multiple psychologies of space. In Frank, A.U. and Campari, I., editors, *Spatial information Theory: A theoretical basis for GIS*, volume 716 of *Lecture Notes in Computer Science*, pages 312–321, Berlin, Heidelberg, New York. Springer Verlag.

Nebel, Bernhard (1997). Solving hard qualitative temporal reasoning problems: Evaluating the efficiency of using the ORD-Horn class. *CONSTRAINTS*, 3(1):175–190.

Nebel, Bernhard and Bürckert, Hans-Jürgen (1995). Reasoning about temporal relations: A maximal tractable subclass of Allen's interval algebra. *Journal of the Association for Computing Machinery*, 42(1):43–66.

Papadias, Dimitris and Theodoridis, Yannis (1997). Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographic Information Systems*, 11(2):111–138.

Papadimitriou, Christos H. (1994). *Computational Complexity*. Addison-Wesley, Reading, MA.

Piaget, Jean and Inhelder, Bärbel (1948). *La représentation de l'espace chez l'enfant*. Presses universitaires de France, Paris, France.

Pratt, Ian and Schoop, Dominik (1998). A complete axiom system for polygonal mereotopology of the real plane. *Journal of Philosophical Logic*, 27: 621–658.

Pujari, Arun K., Kumari, G. Vijaya, and Sattar, Abdul (1999). INDU: An interval and duration network. In *Australian Joint Conference on Artificial Intelligence*, pages 291–303.

Randell, David A. and Cohn, Anthony G. (1989). Modelling topological and metrical properties in physical processes. In Brachman, R., Levesque, H. J., and Reiter, R., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 1st International Conference*, pages 55–66, Toronto, ON. Morgan Kaufmann.

Randell, David A., Cui, Zhan, and Cohn, Anthony G. (1992). A spatial logic based on regions and connection. In Nebel, B., Swartout, W., and Rich, C., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the 3rd International Conference*, pages 165–176, Cambridge, MA. Morgan Kaufmann.

Renz, Jochen (1998). A canonical model of the Region Connection Calculus. In Cohn et al., 1998, pages 330 – 341.

Renz, Jochen (1999). Maximal tractable fragments of the Region Connection Calculus: A complete analysis. In IJCAI-99, 1999, pages 448–454.

Renz, Jochen (2001). A spatial odyssey of the interval algebra: 1.directed intervals. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 51–56, Seattle, WA.

Renz, Jochen (2002). *Qualitative Spatial Reasoning with Topological Information*, volume 2293 of *Lecture Notes in Artificial Intelligence*. Springer Verlag, Berlin, Heidelberg, New York.

Renz, Jochen (2007). Qualitative spatial and temporal reasoning: Efficient algorithms for everyone. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, January 6-12, pages 526–531, Hyderabad, India.

Renz, Jochen and Mitra, Debasis (2004). Qualitative direction calculi with arbitrary granularity. In *PRICAI 2004: Trends in Artificial Intelligence, 8th Pacific Rim International Conference on Artificial Intelligence*, pages 65–74.

Renz, Jochen and Nebel, Bernhard (1999). On the complexity of qualitative spatial reasoning: A maximal tractable fragment of the Region Connection Calculus. *Artificial Intelligence*, 108(1–2):69–123.

Renz, Jochen and Nebel, Bernhard (2001). Efficient methods for qualitative spatial reasoning. *Journal of Artificial Intelligence Research*, 15:289–318.

Schaefer, Thomas J. (1978). The complexity of satisfiability problems. In *Proc. 10th Ann. ACM Symp. on Theory of Computing*, pages 216–226, New York. Association for Computing Machinery.

Schoop, Dominik (1999). *A model-theoretic approach to mereotopology*. PhD thesis, Faculty of Science and Engineering, University of Manchester.

Scivos, Alexander and Nebel, Bernhard (2001). Double-crossing: Decidability and computational complexity of a qualitative calculus for navigation. In *Spatial Information Theory: Foundations of Geographic Information Science*.

Skiadopoulos, Spiros and Koubarakis, Manolis (2005). On the consistency of cardinal direction constraints. *Artificial Intelligence*, 163(1):91–135.

Smith, Terence R. and Park, Keith K. (1992). Algebraic approach to spatial reasoning. *International Journal of Geographic Information Systems*, 6(3): 177–192.

Tarski, Alfred (1941). On the calculus of relations. *Journal of Symbolic Logic*, 6:73–89.

van Beek, Peter (1992). Reasoning about qualitative temporal information. *Artificial Intelligence*, 58(1–3):297–321.

van Beek, Peter and Manchak, Dennis W. (1996). The design and experimental analysis of algorithms for temporal reasoning. *Journal of Artificial Intelligence Research*, 4:1–18.

Vilain, Marc B. and Kautz, Henry A. (1986). Constraint propagation algorithms for temporal reasoning. In *Proceedings of the 5th National Conference of the*

American Association for Artificial Intelligence, pages 377–382, Philadelphia, PA.

Vilain, Marc B., Kautz, Henry A., and van Beek, Peter (1989). Constraint propagation algorithms for temporal reasoning: A revised report. In Weld, D. S. and de Kleer, J., editors, *Readings in Qualitative Reasoning about Physical Systems*, pages 373–381. Morgan Kaufmann, San Mateo, CA.

Whitehead, Alfred N. (1929). *Process and Reality*. The MacMillan Company, New York.

Chapter 5

MODAL LOGICS OF SPACE

Johan van Benthem
University of Amsterdam & Stanford University

Guram Bezhanishvili
New Mexico State University

Second Reader

Marco Aiello
University of Groningen

1. Modal logics and spatial structures

1.1 What does modal logic have to do with space?

Despite historical links between the foundations of mathematics and development of axiomatic geometry, substantial logics for significant spatial structures have been scarce. Perhaps the best-known examples are both due to Tarski. The first is his still amazing work on the first-order theory of elementary Euclidean geometry, including the surprising proof of its decidability, and the resulting abstract theory of real-closed fields. This was the metamathematical finale to Hilbert's *Foundations of Geometry*, itself the culmination of Euclid's *Elements*. This strand is taken up by several chapters in this handbook (Ch. 2, Ch. 7), but it will be only mentioned in passing in this chapter. For our purposes here, the founding event is Tarski's topological interpretation of modal logic, culminating in his proof with McKinsey that the simple decidable modal logic **S4** is complete for interpreting modal $\Diamond$ as topological closure on the reals or any metric space like it. In what follows we concentrate on the latter *modal* direction in spatial logics, which is also represented in several other chapters of the handbook (Ch. 3, Ch. 7, Ch. 10).

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 217–298.
© 2007 *Springer.*

It seems fair to say that there are mostly scattered results in this modal line, suggestive though they may be. To quickly survey several diverse directions in this line we recall Segerberg, 1973 on two-dimensional modal logics, Shehtman, 1983 on logics of physical structures (which was part of Dragalin's program of investigating modal logics of geometrical structures in physical spaces), Goldblatt, 1980 on the logic of Minkowski space-time, Chellas, 1980 on neighborhood semantics (originally proposed by Montague and Scott in the 1960s), the appendix of van Benthem, 1983b on calculi for relative nearness, the work of the 'Georgian School' in modal logics of topology (partly surveyed in Esakia, 2004), Bennett, 1995 on the 'calculus of regions', Venema, 1999 on 'compass logic' in the two-dimensional plane, and Stebletsova, 2000 and Stebletsova and Venema, 2001 on modal logics for projective geometry. So far these ingredients have never added up to one coherent tradition of 'spatial logic', although some attempts have been made occasionally (cf. Anger et al., 1996). In contrast to this state of affair, *temporal logic* has been a thriving research program for many years (cf. van Benthem, 1995 or Hodkinson and Reynolds, 2006). One of the goals of this handbook in general and our chapter in particular is to fill in this gap.

Our starting point is the topological interpretation of modal logic (Tarski, 1938; McKinsey and Tarski, 1944), which we state in the modern truth-conditional format. The basic language $\mathcal{L}$ has a countable set P of proposition letters, boolean connectives $\neg$, $\vee$, $\wedge$, $\rightarrow$, and modal operators $\Box$, $\Diamond$. A *topological model* or simply a *topo-model* is a topological space $\langle X, \tau \rangle$ equipped with a valuation function $\nu : P \rightarrow \mathcal{P}(X)$.

DEFINITION 5.1 (BASIC TOPOLOGICAL SEMANTICS) Truth of modal formulas is defined inductively at points x in a topo-model $M = \langle X, \tau, \nu \rangle$:

$$\begin{array}{lll}
M, x \models p & \text{iff} & x \in \nu(p) \text{ for each } p \in P \\
M, x \models \neg\varphi & \text{iff} & \text{not } M, x \models \varphi \\
M, x \models \varphi \wedge \psi & \text{iff} & M, x \models \varphi \text{ and } M, x \models \psi \\
M, x \models \Box\varphi & \text{iff} & \exists U \in \tau (x \in U \text{ and } \forall y \in U M, y \models \varphi) \\
M, x \models \Diamond\varphi & \text{iff} & \forall U \in \tau (x \in U \rightarrow \exists y \in U : M, y \models \varphi).
\end{array}$$

As usual we can economize by defining, e.g., $\varphi \vee \psi$ as $\neg(\neg\varphi \wedge \neg\psi)$, and $\Diamond\varphi$ as $\neg\Box\neg\varphi$. We will do this whenever convenient.

This usual symbolic truth definition has an immediate spatial interpretation. Given any concrete model, each formula of the language denotes a region of the topological space being modelled. For instance, take the real plane $\mathbb{R}^2$ with the standard topology. Consider a valuation function having some spoon shaped region as the value of the proposition letter p, as depicted in Fig. 5.1.a. Then, the formula $\neg p$ denotes the region not occupied by the spoon, i.e., the background; the formula $\Box p$ denotes the interior of the spoon region p and so on, as explained in Fig. 5.1.

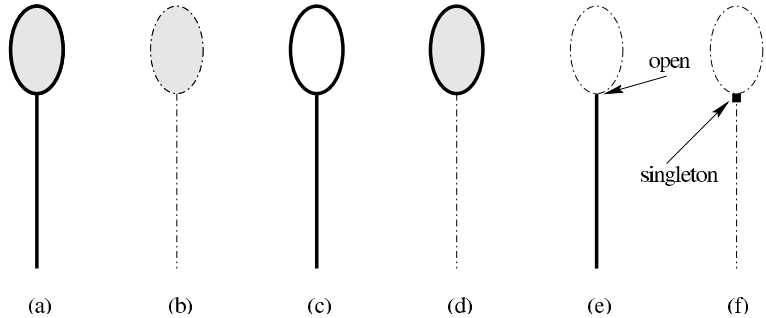


Figure 5.1. Each modal formula identifies a region in a topological space. (a) A spoon, p. (b) The container part of the spoon, $\Box p$. (c) The boundary of the spoon, $\Diamond p \wedge \Diamond \neg p$. (d) The container part of the spoon with its boundary, $\Diamond \Box p$. (e) The handle of the spoon, $p \wedge \neg \Diamond \Box p$. In this case the handle does not contain the junction handle-container point. (f) The junction handle-container point of the spoon, $\Diamond \Box p \wedge \Diamond (p \wedge \neg \Diamond \Box p)$: a singleton in the topological space.

Thus, a simple modal language can define regions in space in a perspicuous and appealing way, and allow us to check assertions about them. Moreover, the same modal notation also facilitates spatial reasoning. For instance, the valid axiom $\Box(p \wedge q) \leftrightarrow (\Box p \wedge \Box q)$ says that two ways of describing a region—either as the interior of the intersection of two sets or as the intersection of the interiors of those sets—always amount to the same thing. Thus, modal logic is also a small inference engine for basic spatial manipulations.

We will consider other modal languages and logics for spatial structures later on. For the moment we merely point out that the preceding example contains two different perspectives on the encounter between modal logic and space. Some modal logicians see topological models as a means of providing *new semantics* for existing modal languages, mostly for logic-internal purposes. This can be motivated a bit more profoundly by thinking of topologies as models for *information*, making this interest close to central logical concerns. But someone primarily interested in space as such will not worry about the semantics of modal languages. She will rather be interested in spatial structures by themselves, and *spatial logics* will be judged by how well they analyze old structures, discover new ones, and help in reasoning about them. Both perspectives will play in our presentation, with the mathematics largely the same, but the sort of issues suggested sometimes a bit different.

In the remainder of this introductory section we discuss these issues in more detail, as a first pass through the topic. In Sec. 1.7 we then describe the setup for the rest of this chapter.

1.2 Relational semantics for modal logic

The standard models for modal logic are the well-known binary relational graphs, with necessity interpreted as truth in all accessible worlds, and possibility as truth in at least one accessible world:

$$M, s \models \Box\varphi \text{ iff } \forall t(sRt \rightarrow M, t \models \varphi)$$
$$M, s \models \Diamond\varphi \text{ iff } \exists t(sRt \text{ and } M, t \models \varphi).$$

In this chapter we presuppose a basic acquaintance with modal logic in this style. We refer to Blackburn et al., 2001 and van Benthem and Blackburn, 2006 for a quick introduction in a modern spirit. In particular, here are some core themes that will occur below.

The natural measure of expressive power for the basic modal language over the class of arbitrary relational models is the invariance of all formulas under *bisimulations* between models M, w and N, v, which provides the right measure for structural equivalence as far as the language is concerned. This invariance analysis can be fine-tuned to play Ehrenfeucht-Fraisse-type model comparison games between models (Doets, 1996) in which the Duplicator player has a winning strategy over a k-round game iff the two models M, w and N, v satisfy the same modal formulas up to modal operator depth k. As for axiomatics, the class of all standard models validates precisely the *minimal modal logic* **K**, whose most noteworthy principle is the above-mentioned distributivity of modal $\Box$ over conjunction. But deductive power goes up on special model classes. E.g., the modal logic **S4** with axioms $\Box p \rightarrow p$ and $\Box p \rightarrow \Box\Box p$ is complete for the class of all reflexive and transitive frames, and there is a host of other natural stronger logics. These *correspondences* between natural conditions on accessibility relations in graphs and modal axioms of certain shapes can also be studied per se as a matter of semantic definability. There are even powerful methods for automatic analysis of modal axioms for their frame content. But in addition to deductive power and correspondence analysis of the basic language, there is also expressive power: the ability to say more about the same class of structures. Many modal languages in use *extend* the basic propositional formalism mentioned above by adding operators such as the 'universal modality' ('true in all worlds'), or temporal-style operators like 'Until' or 'Since'.

Finally, to complete this lightning summary, modal languages are designed with a certain *balance* in mind. E.g., the basic modal language is like the language of first-order logic in that it allows for quantification over objects. But this quantification is only 'local' or 'bounded', tied by accessibility to the current world. Trading in some first-order expressive power in this way, however, comes with a bonus: validity and satisfiability in the basic modal language are *decidable*, indeed *PSPACE*-complete. Moreover, looking at other

key tasks for a logical calculus, it may be noted that *model checking* for finite models is *PSPACE*-complete for the full first-order language, whereas it takes only polynomial time for modal logic. Likewise, testing two finite models for the existence of a bisimulation can be done in polynomial time, whereas the corresponding problem for the complete first-order language is the so-called Graph Isomorphism Problem, which is known to be in *NP*. More generally, extended modal languages try to boost expressive power on relevant structures, while skirting the cliffs of complexity. Well-known examples of such trade-offs much higher up are the 'guarded fragment' of first-order logic (Andréka et al., 1998) or the non-first-order modal 'μ-calculus' enriching the basic modal language with non-first-order operators for smallest and greatest fixed-points (Harel et al., 2000).

Even though these features of modal logic have not evolved for specific spatial reasons, they are often germane to thinking about space. First of all, binary relational models themselves *are* a form of geometrization of modal semantics. Of course, they resemble abstract graphs and diagrams rather than regions of Euclidean spaces, but still, geometrical intuitions play a role in understanding how it all works. Indeed, models of this sort can represent significant spatial structures. An example is the work of Shehtman, 1983 and Goldblatt, 1980 from the early 1980s (cf. also Ch. 11). Interestingly, in relativistic space-time, the crucial primitive notion is not the ternary 'betweenness' of classical geometry, but the binary relation Cxy of *forward causal accessibility*, which runs from a point x to all points y in the interior of its future light-cone, where causal signals can reach (see Fig. 5.2.a).

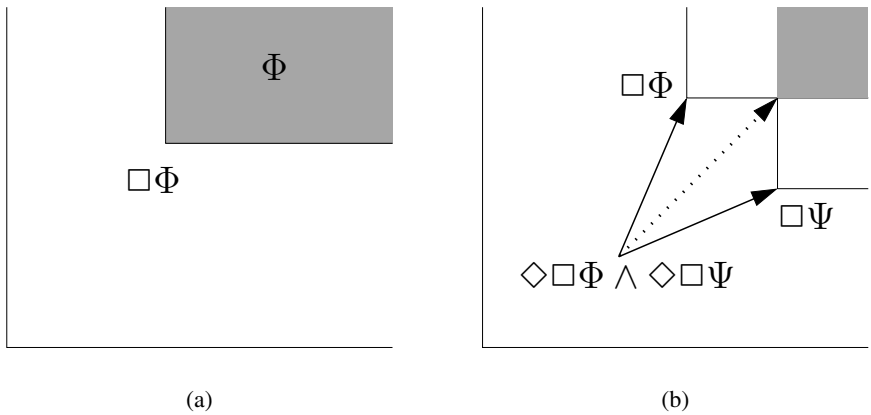


Figure 5.2. Forward modality in Minkowski space-time and validity of the **S4.2** Confluence Axiom.

Shehtman and Goldblatt independently proved that the complete modal logic of forward causal accessibility equals the modal logic **S4.2** which extends **S4**

with the so-called 'Confluence Axiom' $\Diamond\Box p \to \Box\Diamond p$. The latter principle is illustrated in Fig. 5.2.b. It expresses the relativistic fact that any two different causal futures, as seen from the current point, even when not causally connected themselves, could potentially still lead to a common future history. Again we see how modal formulas express significant facts about space(–time).

All technical topics in our survey of relational semantics make spatial sense. Bisimulation-invariance analysis of expressive power is very close to thinking about geometrical *transformations* and *invariants* (van Benthem, 2002), which goes back to the foundations of geometry in the 19th century. Also, modal logics can represent special styles of spatial reasoning, as we just saw. And issues of optimal language design have also emerged already. For instance, the above topological semantics for the basic modal language is still 'local', not in the sense of binary accessibility, but in being restricted to what is true in open neighborhoods of the current point. But many natural topological notions do not have this local character. E.g., a space is *connected* if it cannot be split into two non-empty clopen sets. This global property of topological spaces cannot be expressed in the basic modal language. But it can if one adds a universal modality. Finally, in all this, the issue of 'balance' returns. Modal systems are typically attempts at uncovering significant spatial structures, while providing low-complexity (decidable) calculi for reasoning with them.

1.3 Background: the many semantics of modal logic

In a sense, spatial interpretations of modal logic challenge the existing order. The now dominant relational semantics is really a product of the 1950's/1960's. Its historical predecessors include *algebraic semantics*, which has been used extensively in the technical literature in the form of boolean algebras with operators. Venema, 2006 surveys the state of the art. Another earlier semantics of modality is Gödel's provability interpretation: a story which is told with many new historical details in Artemov, 2006. The latter paper is also an excellent broader source for mathematical uses of modal logic, including a brief, but useful account of spatial ones.

Clearly, our topological semantics is another 1930's challenger. This modelling was particularly vivid and attractive for the language of *intuitionistic logic*, where open sets may be viewed as information stages concerning some underlying point—an interpretation which returns in much greater sophistication in the topos semantics (see Ch. 8). The informational interpretation of topology will not be a major concern in this chapter, but we do mention topological semantics for *epistemic logic* briefly in Sec. 3.4 as it raises some interesting new issues that do not become visible in the standard binary relational modelling.

Also worth noting is that the topological semantics generalizes easily to the so-called *neighborhood models* for modal languages. Here one just assumes some binary relation RxY associating worlds x with sets of worlds Y (not necessarily open environments), with the same truth condition as above:

$\Box\varphi$ is true at world x iff there exists a set Y with RxY
all of whose members y satisfy the formula φ.

Neighborhood models are used, e.g., to express output relations for concurrent computation (Peleg, 1987), relations of 'support' or 'dependence' in logic programming (van Benthem, 1992), or relations of 'power' for forcing a game to end in certain sets of outcomes in games, starting from some current node (Pauly, 2001). Neighborhood semantics is an interesting counterpoint to topological semantics, because it shows what happens further down the road. The minimal modal logic now loses Distributivity, retaining only *upward monotonicity* for the two modalities. There is still a notion of generalized bisimulation, however, whose topological version will return in Sec. 1.4. Finally, as to the balance, the complexity of satisfiability in neighborhood semantics goes down from *PSPACE-complete* to *NP*. The latter is not true, however, for the topological interpretation, as it retains the Distribution Axiom, and its minimal modal logic **S4** is still *PSPACE-complete*.

All these different semantics are related. In particular, topological models are a special case of neighborhood models, and reflexive and transitive relational models are a special case of topological models, as will be explained below. Neighborhood models are also related to algebraic ones, but we will forego such details in this chapter. Even so, these technical connections have their uses. For instance, topological semantics still includes binary relational semantics as the special case of 'Alexandroff topologies' (cf. Sec. 2.4.1). Thus, its generalizations of standard modal notions, such as bisimulations, may be viewed as a significant extension of the latter's scope of applicability. Likewise, we will see in Sec. 3.2 how a topological perspective actually clarifies issues in binary relational model theory, viz. the axiomatization problem for classes of products of modal frames. And finally, topological viewpoints have suggested new modal languages and structures such as the 'Chu spaces' of Pratt, 1999 (cf. van Benthem, 2000a on a first-order/modal style analysis of invariance and expressive power).

We conclude with one illustration going the other way. Despite its immediate spatial appeal, the topological semantics is also *more complex* than the binary relational semantics. Instead of matching up one modal operator $\Box$ with one quantifier, it matches it up with the $\exists\forall$ combination of *two* nested quantifiers: 'there exists an open set such that for all its elements...'. This makes things less perspicuous, and it may in fact be the reason why the topological interpretation, though historically first, was eventually supplanted by the simpler

Kanger-Hintikka-Kripke graph-based version. But this is not all there is to be said. For, one can analyze the above $\exists\forall$ in terms of two consecutive modalities $\Diamond_{open}\Box_{element}$, where the first states the existence of an open set, while the second accesses its elements. From this point of view, the topological semantics lives inside a standard *bimodal* language over two-sorted binary relational models having both points and sets as objects. There are even mathematical reduction results showing precisely how far this reduction goes. This amounts to a richer many-sorted view of space, where both points and sets can be 'objects' on a par. This style of thinking, too, is in the line with the geometrical tradition, which has points, lines, and spaces as objects on a par rather than ascending stages in some abstract set-theoretic hierarchy. van Benthem, 1999 presents a defense of many-sorted reformulations of complex modal semantics in temporal and spatial settings. The only framework that we know of where this 'unravelling' into separate modal stages is taken seriously in a spatial sense is the 'topological logic of knowledge' of Dabrowski et al., 1996 (cf. also Ch. 6). The bulk of existing work, however, is squarely within the standard topological framework, to which we now return.

1.4 Modal logic and topology: first steps

The topological interpretation explained above brings some interesting shifts in perspective. E.g., the crucial modal feature of *locality* in graph models now means that a formula is true at M, x iff it is true at x in any submodel whose domain is that of M restricted to some *open neighborhood* of x. Thus, *regions* are essential, and more generally, a modal approach provides a calculus of regions de-emphasizing constellations of points. As such, it is close to 'region versus points' theories of time and space (Allen, 1983; Allen and Hayes, 1985; van Benthem, 1983b; Randell et al., 1992).

There are also subtle differences with modal logic that lie just below the surface. E.g., binary relational semantics validates unlimited *distributivity*: the modal box distributes over arbitrary infinite conjunctions of formulas. This is not so in topological semantics:

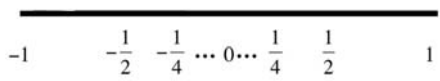


Figure 5.3. Nested intervals refuting countable distributivity.

Let the proposition letters p_i be interpreted as the open intervals $(-1/i, 1/i)$. Then the point 0 satisfies the countable conjunction of all formulas $\Box p_i$. But 0 does not satisfy the related infinitary modal formula with the box over the

conjunction, since that intersection is just the set $\{0\}$, whose topological interior is empty (see Fig. 5.3).

1.4.1 Expressive power: topo-bisimulation and topo-games. To understand the expressive power of a modal language, a suitable notion of bisimulation is needed. The following definition reflects the semantic definition of the modal operators and can be seen as composed of two sub-moves: one in which points are linked, and one in which containing opens are matched.

DEFINITION 5.2 (TOPO-BISIMULATION) A *topological bisimulation* or simply a *topo-bisimulation* between two topo-models $M = \langle X, \tau, \nu \rangle$ and $M' = \langle X', \tau', \nu' \rangle$ is a non-empty relation $T \subseteq X \times X'$ such that if xTx' then:

1 $x \in \nu(p) \Leftrightarrow x' \in \nu'(p)$ for each $p \in P$

2 (forth): $x \in U \in \tau \rightarrow \exists U' \in \tau' : x' \in U'$ and $\forall y' \in U' \; \exists y \in U : yTy'$

3 (back): $x' \in U' \in \tau' \rightarrow \exists U \in \tau : x \in U$ and $\forall y \in U \; \exists y' \in U' : yTy'$.

A topo-bisimulation is *total* if its domain is X and its range is X'. If only the atomic clause (1) and the forth condition (2) hold, we say that the second model *simulates* the first.

REMARK 5.3 We point out that the forth condition in the definition of topo-bisimulation is equivalent to the T-image of every open subset of $\langle X, \tau \rangle$ being open, while the back condition is equivalent to the T-inverse image of every open subset of $\langle X', \tau' \rangle$ being open.

Topo-bisimulation captures the adequate notion of 'model equivalence' for the basic language $\mathcal{L}$ topologically interpreted. Evidence for this comes from the following two results (cf. Aiello and van Benthem, 2002a).

THEOREM 5.4 *Let $M = \langle X, \tau, \nu \rangle$ and $M' = \langle X', \tau', \nu' \rangle$ be two topo-models, and $x \in X$ and $x' \in X'$ be two topo-bisimilar points. Then for each modal formula φ we have $M, x \models \varphi$ iff $M', x' \models \varphi$. That is, modal formulas are invariant under topo-bisimulations.*

THEOREM 5.5 *Let M, M' be two finite models, and $x \in X$, $x' \in X'$ be such that for each φ we have $M, x \models \varphi$ iff $M', x' \models \varphi$. Then there exists a topo-bisimulation between M and M' connecting x and x'. That is, finite modally equivalent models are topo-bisimilar.*

Topo-bisimulations are coarsenings of the basic structural equivalence in topology:

THEOREM 5.6 *If two topological spaces $\langle X, \tau\rangle$ and $\langle X', \tau'\rangle$ are homeomorphic, then for each valuation ν on $\langle X, \tau\rangle$ there exists a valuation ν' on $\langle X', \tau'\rangle$ such that the topo-models $M = \langle X, \tau, \nu\rangle$ and $M' = \langle X', \tau', \nu'\rangle$ are topo-bisimular.*

See Aiello and van Benthem, 2002a for details as well as connections with other topological notions of structural similarity such as homotopy.

Topo-bisimulation is a standard model-theoretic tool for assessing expressivity of our language with respect to spatial patterns. Nevertheless, when comparing e.g. two image representations, it may still be too coarse. To refine the similarity matching, one can define a topological *model comparison game* $TG(M, M', n)$ between two topo-models M, M'. The idea of the game is that two players challenge each other picking elements from the two models to compare. 'Spoiler' wins if he can show the models to be different, 'Duplicator' wins if he can show the models to be 'similar'. Winning strategies for Duplicator in infinite games, requiring never-ending continued responses, match up precisely with topo-bisimulations. Furthermore, for finite-length games, games and modal formulas are connected by the Adequacy Theorem:

THEOREM 5.7 (AIELLO AND VAN BENTHEM, 2002A) *Duplicator has a winning strategy in the topo-game $TG(M, M', n, x, x')$ iff x and x' satisfy the same formulas of modal operator depth up to n in their respective models M, M'.*

Fig. 5.4 shows how many rounds Spoiler will need to distinguish positions on cutlery. The number of rounds corresponds to the depth of a modal 'difference formula' for the points under comparison. For instance, the single round in 2(a) corresponds to $\Box p$ versus $\neg\Box p$, the two rounds in 2(b) correspond to $\neg\Diamond\Box p$ versus $\Diamond\Box p$, while the three rounds in 2(c) correspond to $\Diamond(p \wedge \neg\Diamond\Box p)$ versus $\neg\Diamond(p \wedge \neg\Diamond\Box p)$. The formal definition of a game, and an extensive discussion of plays and strategies are in Aiello and van Benthem, 2002a, while algorithms to compute winning strategies for comparing models are illustrated in Aiello, 2002b.

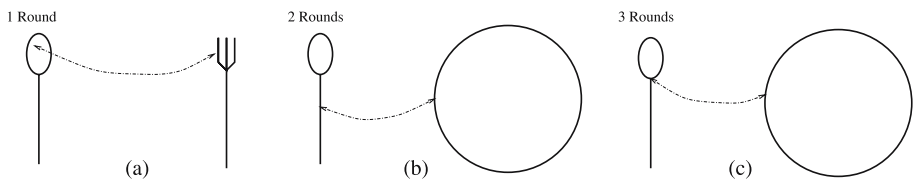


Figure 5.4. Game rounds needed for distinguishing shapes.

Excursion: Our examples show how logical games match topological notions very well. But there is a much earlier historical precedent. It is shown in van Dalen, 2005 how Brouwer defined the crucial topological notion of *dimension* in terms of the following game:

Player 1 chooses two disjoint closed subsets A_1, B_1 of the space. Player 2 then chooses a closed separating set S_1. Player 1 now chooses two disjoint closed subsets of S_1. Player 2 then chooses a closed separating set S_2 inside S_1, etc.

Player 2 wins if a separating set S_n is reached after n rounds which is totally disconnected. According to Brouwer, the dimension of a space is the lowest natural number n for which Player 2 has a winning strategy in the n-round game.

1.4.2 Deductive power: topo-logics. Now consider logical validity and hence the general calculus for spatial reasoning in this language. The logic **S4** is defined by the KT4 axioms and the rules of Modus Ponens and Necessitation (see Sec. 2.2 below). In the topological setting these principles translate into the following ones, with an informal explanation added:

$\Box\top$	(N)	the whole space is open
$(\Box p \wedge \Box q) \leftrightarrow \Box(p \wedge q)$	(R)	open sets are closed under finite intersections
$\Box p \rightarrow \Box\Box p$	(4)	the interior operator is idempotent
$\Box p \rightarrow p$	(T)	the interior of any set is contained in the set.

Then the universally valid formulas topologically interpreted are precisely the theorems of **S4**. But McKinsey and Tarski, 1944 proved a much more striking result.

THEOREM 5.8 **S4** *is complete for any dense-in-itself metric separable space.*

Thus, **S4** is also the logic of any Euclidean space $\mathbb{R}^n$ with the standard topology. Mints, 1998 proved completeness of **S4** for the Cantor space in a particularly elegant manner.

More restricted spatial structures generate stronger modal logics on top of **S4**. Take, for instance, the *serial* subsets of the real line, being the finite unions of convex intervals (Aiello et al., 2003). These have been used to model life-spans of 'events' in linguistics and computer science. Now consider the following additional axioms:

(BD$_2$) $(\neg p \wedge \Diamond p) \rightarrow \Diamond\Box p$

(BW$_2$) $\neg(p \wedge q \ \wedge \ \Diamond(p \wedge \neg q) \ \wedge \ \Diamond(\neg p \wedge q) \ \wedge \ \Diamond(\neg p \wedge \neg q))$.

These are complete for the serial sets. To give an impression of what is going on, look at Fig. 5.5, with a serial set denoted by p, and take the axiom **BD$_2$**.

```
p          ..... ( ________ ] ..... x ......... ( _____ )( _____ ] .........
                   0        √2    2        3        4        5

¬p ∧ ◇p    ..... x ......... ... ................... x ......... x ..........

◇□p        ..... [ ________ ] ................... [ _____________ ] .........
```

Figure 5.5. A serial set of $\mathbb{R}$ and the defined subformulas by the axiom $\mathbf{BD}_2$.

In relational semantics, this axiom bounds the depth of the model by 2. In topological semantics, it states that the points that are both in the complement $\neg p$ of a region and in its closure $\Diamond p$, must be in the regular closed portion $\Diamond\Box p$ of the region itself.

Similarly, one can look at interesting 2-dimensional topological spaces. Here is a modal axiom

$$(\mathbf{BD}_3) \qquad \Diamond(\Box p_3 \wedge \Diamond(\Box p_2 \wedge \Diamond\Box p_1 \wedge \neg p_1) \wedge \neg p_2) \rightarrow p_3$$

valid in the 'rectangular serial' sets of the plane $\mathbb{R}^2$. These special structures are investigated in Aiello et al., 2003 and van Benthem et al., 2003. The latter provides an axiomatization for logics of this sort for Euclidean spaces of any dimension (see Sec. 2.6 below).

1.5 Modal logics of other spatial structures

Our account so far may have suggested that modal logic of space *must* be about topology. But this is not the case at all. Moving on from topology to more 'rigid' spatial structures, modal logic returns just as well, though in new guises. For instance, consider *affine geometry*, where the major notion is a ternary notion of *betweenness* $\beta(xyz)$, which says that point y lies in between points x and z, allowing y to be one of these endpoints. Now define a binary betweenness modality $<B>$:

$$M, x \models <B>(\varphi, \psi) \quad \text{iff} \quad \exists y, z : \beta(yxz) \text{ and } M, y \models \varphi \text{ and } M, z \models \psi.$$

Again, this leads to very concrete spatial pictures. This time standard geometrical figures can be described by modal means. Consider Fig. 5.6. Let the proposition letter p denote the set of three points on the left forming the vertices of a triangle. The next two phases of the picture show how the formula $<B>(p, p)$ holds on the sides of the triangle, while the whole triangle, including its interior, is defined by the modal formula $<B>(<B>(p, p), p)$.

Clearly all of the earlier technical modal notions make sense once more. For instance, we can study modal bisimulation between geometric figures, or modal deduction about triangles or convex figures. We will study such issues

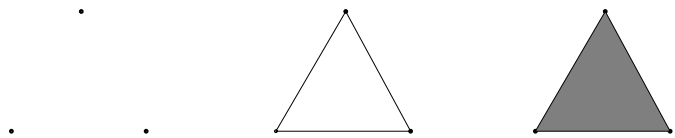


Figure 5.6. The formulas p, $<B>(p,p)$ and $<B>(<B>(p,p),p)$.

for geometrical modal logics later in Sec. 4. For the moment, we just note how this formalism can express significant geometric facts in surprising ways. Consider the following basic geometrical principle, known as '*Pasch's Axiom*' (see Fig. 5.7), written as follows in first-order notation:

$$\forall txyzu(\beta(xtu) \;\&\; \beta(yuz) \rightarrow \exists v : \beta(xvy) \;\&\; \beta(vtz))$$

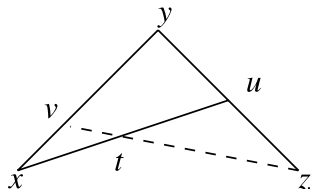


Figure 5.7. Pasch's property.

It says that any line drawn through a vertex of a triangle and continuing into its interior must cross the opposite side to that vertex at some point. This does not look modal at all, but in fact it is! Consider the following axiom of *associativity* for the betweenness modality:

$$<B>(p, <B>(q,r)) \rightarrow <B>(<B>(p,q),r).$$

We will see in Sec. 4 that:

FACT 5.9 *Modal Associativity corresponds to Pasch's Axiom.*

At this stage a useful exercise for the reader would be to check how, unpacking the nested modalities $<B>$ in the antecedent, Pasch's Axiom is in fact precisely what is needed to see the validity of Associativity. In Sec. 4 and 5 we will develop these ideas further to also include metric geometry and eventually even linear algebra.

1.6 Logical analysis of space once more

Let us summarize the methodology of this chapter once more. Topologists or geometers do not worry about formal systems: they just state what they

see in whatever formalism at hand. Logicians, however, propose a trade-off: specify a formal language restricting the notions one can talk about, and then see what complete logic comes out, perhaps even in the form of a decidable calculus. Tarski's elementary geometry is still a paradigm for this approach, and so are other logics discussed in Ch. 2 of this handbook on first-order theories of polygons and of mereotopology. Incidentally, the spatial perspective also highlights quite different uses of a logical formalism. One is its *descriptive role* in defining spatial patterns, allowing us to describe these, check whether they hold in given situations, and compare different properties of spatial structures. Another is its *deductive role* as a calculus of reasoning about space, which is associated with other tasks, such as mathematical theorizing, information extraction from spatial databases, or reasoning by a robot trying to plan actions in a partially unknown environment.

Used in either mode, modal languages are fragments of first-order ones, restricting expressive power even further, but promising better complexity. The very multiplicity of modal languages is an advantage here, as we can work at different levels of structure, measured by different notions of invariance, whether topo-bisimulation, or some logical or geometrical strengthening thereof. This fits with the mathematical idea that space can be studied legitimately at various levels of detail. Finally, consider the issue of the modal balance between expressive power and computational complexity. Indeed, there are low-complexity modal logics for some spatial structures. But there are also some phenomena showing that things can be complicated. For instance, in affine geometry, while the minimal modal logic of our binary betweenness modality $<B>(\varphi, \psi)$ is decidable, this same language becomes undecidable over the special class of associative relations that we just associated with Pash's Axiom. The reason is that it can then straightforwardly encode the word problem for semigroups.

Moreover, our two guiding examples from Tarski's work do not point in the same direction in terms of the sources of their decidability. Modal Topology is indeed decidable for reasons of modal parsimony by abstract general methods having little to do with peculiarities of topological spaces. But Elementary Geometry is decidable not because its language has judiciously toned-down expressive power, but because its intended model of Euclidean Space is so rich that it happens to support a procedure of *quantifier elimination*, providing us with the decision algorithm, which is very special for this geometric setting. We refer to Ch. 9 for more accumulated evidence on complexity of logics for spatial reasoning.

1.7 Contents of this chapter

This concludes our introduction to modal languages of spatial structures. The rest of this chapter is organized as follows. Sec. 2 contains a more extensive

formal treatment of topological models, leading up to the general comple teness theorem for topological models, and from there, to the landmark completeness theorem for the reals using modern techniques. Following that, we also discuss richer modal logics for more special topological structures, including placing restrictions on sets that can serve as values for propositions. Next, Sec. 3 surveys a number of more recent special topics in this area. These include (a) alternative interpretations of the modalities in terms of the topological derivative, (b) combining modal logics for describing products of topological spaces, (c) language extensions that can express further topological structure, and finally (d) topological models for epistemic logic, with an excursion into fixed-point extensions of the language that can define various notions of common knowledge. Sec. 4 is a discussion of modal languages for geometric structures, starting with affine cases, and then moving to modal languages for metric relations of relative nearness. We also provide a comparison with first-order languages for these structures. Finally, Sec. 5 looks at modal logics for mathematical morphology, which basically amounts to analyzing certain subsets of vector spaces. This topic also involves connections with modal 'arrow logics' for analyzing structures in relational algebra. Sec. 6 contains our conclusions.

2. Modal logic and topology: basic results

2.1 Topological preliminaries

We start by surveying briefly the basic topological concepts that will be used throughout this chapter. They can be found in any textbook on general topology (see, e.g., Engelking, 1989; Kelley, 1975; Kuratowski, 1966).

DEFINITION 5.10 *A* topological space *is a pair* $\mathcal{X} = \langle X, \tau \rangle$*, where* X *is a nonempty set and* τ *is a collection of subsets of* X *satisfying the following three conditions:*

1 $\emptyset, X \in \tau$;

2 *If* $U, V \in \tau$*, then* $U \cap V \in \tau$;

3 *If* $\{U_i\}_{i \in I} \in \tau$*, then* $\bigcup_{i \in I} U_i \in \tau$.

The elements of τ are called *open sets*. The complements of open sets are called *closed sets*. An open set containing $x \in X$ is called an *open neighborhood* of x.

A family $\mathcal{B} \subseteq \tau$ is called a *basis* for the topology if every open set can be represented as the union of elements of a subfamily of $\mathcal{B}$. It is well-known that a family $\mathcal{B}$ of subsets of X is a basis for some topology on X iff (i) for each $x \in X$ there exists $U \in \mathcal{B}$ such that $x \in U$, and (ii) for each $U, V \in \mathcal{B}$, if $x \in U \cap V$, then there exists $W \in \mathcal{B}$ such that $x \in W \subseteq U \cap V$.

For $A \subseteq X$, a point $x \in X$ is called an *interior point* of A if there is an open neighborhood U of x such that $U \subseteq A$. Let $\mathrm{Int}(A)$ denote the set of interior points of A. Then it is easy to see that $\mathrm{Int}(A)$ is the greatest open set contained in A, called the *interior* of A. A point $x \in X$ is called a *limit point* of $A \subseteq X$ if for each open neighborhood U of x, the set $A \cap (U - \{x\})$ is nonempty. The set of limit points of A is called the *derivative* of A and is denoted by $d(A)$. Let $\mathrm{Cl}(A) = A \cup d(A)$. Then it is easy to see that $x \in \mathrm{Cl}(A)$ iff $U \cap A$ is nonempty for each open neighborhood U of x, and that $\mathrm{Cl}(A)$ is the least closed set containing A, called the *closure* of A.

Let Int and Cl denote the interior and closure operators of $\mathcal{X}$, respectively. Then it is well known that the following are satisfied for each $A, B \subseteq X$:

$$\begin{array}{ll} \mathrm{Int}(X) = X & \mathrm{Cl}(\emptyset) = \emptyset \\ \mathrm{Int}(A \cap B) = \mathrm{Int}(A) \cap \mathrm{Int}(B) & \mathrm{Cl}(A \cup B) = \mathrm{Cl}(A) \cup \mathrm{Cl}(B) \\ \mathrm{Int}(A) \subseteq A & A \subseteq \mathrm{Cl}(A) \\ \mathrm{Int}(A) \subseteq \mathrm{Int}(\mathrm{Int}(A)) & \mathrm{Cl}(\mathrm{Cl}(A)) \subseteq \mathrm{Cl}(A). \end{array}$$

Moreover, there is a duality $\mathrm{Int}(A) = X - \mathrm{Cl}(X - A)$, and a topological space can also be defined in terms of an interior operator or a closure operator satisfying the above four conditions.

We also let $t(A)$ denote $X - d(X - A)$. Then $x \in t(A)$ iff there exists an open neighborhood U of x such that $U \subseteq A \cup \{x\}$. We call $t(A)$ the *co-derivative* of A. Let d and t denote the derivative and co-derivative operators of $\mathcal{X}$, respectively. Then it is well known that the following are satisfied for each $A, B \subseteq X$:

$$\begin{array}{ll} d(A \cup B) = d(A) \cup d(B) & t(A \cap B) = t(A) \cap t(B) \\ d(d(A)) \subseteq A \cup d(A) & A \cap t(A) \subseteq t(t(A)). \end{array}$$

DEFINITION 5.11 *Let $\mathcal{X}$ be a topological space and A be a subset of X.*

1 *A is called* clopen *if it is both closed and open.*

2 *A is called* dense *if* $\mathrm{Cl}(A) = X$.

3 *A is called* boundary *if* $\mathrm{Int}(A) = \emptyset$.

4 *A is called* dense-in-itself *if* $A \subseteq d(A)$.

For a topological space $\mathcal{X}$, the family $\{U_i\}_{i \in I} \subseteq \tau$ is called an *open cover* of X if $\bigcup_{i \in I} U_i = X$.

DEFINITION 5.12 *Let $\mathcal{X}$ be a topological space.*

1 *$\mathcal{X}$ is called* discrete *if every subset of X is open.*

2 *$\mathcal{X}$ is called* trivial *if $\emptyset$ and X are the only open subsets of X.*

3 *$\mathcal{X}$ is called* dense-in-itself *if* $d(X) = X$.

4 $\mathcal{X}$ *is called* separable *if there exists a countable dense subset of* X.

5 $\mathcal{X}$ *is called* compact *if every open cover of* X *has a finite subcover.*

6 $\mathcal{X}$ *is called* connected *if* $\emptyset$ *and* X *are the only clopen subsets of* X.

7 $\mathcal{X}$ *is called* 0-dimensional *if clopen subsets of* X *form a basis for the topology.*

8 $\mathcal{X}$ *is called* extremally disconnected *if the closure of each open subset of* X *is clopen.*

In the next definition we recall the separation axioms T_0, T_d, T_1, and T_2.

DEFINITION 5.13 *Let* $\mathcal{X}$ *be a topological space.*

1 $\mathcal{X}$ *is called a* T_0-space *if for each pair of different points there exists an open set containing one and not containing the other.*

2 $\mathcal{X}$ *is called a* T_d-space *if for each* $x \in X$ *there exists an open neighborhood* U *of* x *such that* $\{x\}$ *is closed in* U. *Equivalently,* $\mathcal{X}$ *is a* T_d*-space iff* $dd(A) \subseteq d(A)$.

3 $\mathcal{X}$ *is called a* T_1-space *if for each pair of different points there exists an open set containing exactly one of the points. Equivalently,* $\mathcal{X}$ *is a* T_1*-space iff each* $\{x\}$ *is closed in* X.

4 $\mathcal{X}$ *is called a* T_2-space *or a* Hausdorff space *if for each pair* $x, y \in X$ *of different points there exit disjoint open neighborhoods of* x *and* y.

It is well known that every T_2-space is a T_1-space, that every T_1-space is a T_d-space, and that every T_d-space is a T_0-space, but not vice versa.

Let $\mathcal{X}$ and $\mathcal{Y}$ be topological spaces. We call $\mathcal{Y}$ a *subspace* of $\mathcal{X}$ if $Y \subseteq X$ and U is an open subset of Y iff there exists an open subset V of X such that $U = V \cap Y$.

DEFINITION 5.14 *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be topological spaces and* $f : X \to Y$ *be a map.*

1 f *is called* continuous *if* U *open in* Y *implies* $f^{-1}(U)$ *is open in* X.

2 f *is called* open *if* U *open in* X *implies* $f(U)$ *is open in* Y.

3 f *is called* interior *if it is both continuous and open.*

We call $\mathcal{Y}$ a *continuous image* of $\mathcal{X}$ if there exists a continuous map from X onto Y. Open and interior images of $\mathcal{X}$ are defined analogously.

Let $\{\mathcal{X}_i\}_{i \in I}$ be a family of pairwise disjoint topological spaces. We define the *topological sum* of $\{\mathcal{X}_i\}_{i \in I}$ as the pair $\bigoplus_{i \in I} \mathcal{X}_i = \langle \bigcup_{i \in I} X_i, \tau \rangle$, where $U \in \tau$ iff $U \cap X_i \in \tau_i$. If the members of the family $\{\mathcal{X}_i\}_{i \in I}$ are not pairwise

disjoint, then the topological sum is defined using disjoint union instead of set-theoretic union.

2.2 Relational semantics and some modal logics

Before exploring topological interpretations in more depth, we recall some basic facts from relational semantics.

2.2.1 The uni-modal case. We recall that a *frame* is a relational structure $\mathfrak{F} = \langle W, R\rangle$ such that W is a nonempty set and R is a binary relation on W. A *valuation* of the basic modal language $\mathcal{L}$ in $\mathfrak{F}$ is a function ν from the set P of propositional letters of $\mathcal{L}$ to the powerset of W. A pair $M = \langle \mathfrak{F}, \nu\rangle$ is called a *model* (based on $\mathfrak{F}$). Given a model M, we define when a formula φ is *true at a point* $w \in W$ by induction on the length of φ:

- $w \models p$ iff $w \in \nu(p)$;
- $w \models \neg\varphi$ iff not $w \models \varphi$;
- $w \models \varphi \wedge \psi$ iff $w \models \varphi$ and $w \models \psi$;
- $w \models \Box\varphi$ iff $(\forall v \in W)(wRv \to v \models \varphi)$;

and hence also

- $w \models \Diamond\varphi$ iff $(\exists v \in W)(wRv \;\&\; v \models \varphi)$.

We say that φ is *true* in M if φ is true at every point in W, and that φ is *valid* in $\mathfrak{F}$ if φ is true in every model M based on $\mathfrak{F}$. Finally, we say that φ is *valid* in a class of frames if φ is valid in every member of the class.

Below we list several standard modal logics and their axiomatizations.

DEFINITION 5.15

1 *The basic logic* **K** *of all frames is axiomatized by the axiom:*

(K) $$\Box(p \to q) \to (\Box p \to \Box q).$$

with Modus Ponens and Necessitation as the only rules of inference:

(MP) $$\frac{\varphi \qquad \varphi \to \psi}{\psi}$$ (N) $$\frac{\varphi}{\Box\varphi}$$

2 *The logic* **T** *of reflexive frames is axiomatized by adding to* **K** *the axiom:*

(T) $$\Box p \to p$$

3 *The logic* **K4** *of transitive frames is axiomatized by adding to* **K** *the axiom:*

(4) $$\Box p \to \Box\Box p$$

4 *The logic* **S4** *of reflexive and transitive frames is axiomatized by adding to* **K** *the axioms* (T) *and* (4).

5 *The logic* **S5** *of reflexive, transitive, and symmetric frames is axiomatized by adding to* **S4** *the axiom:*

(B) $$p \to \Box\Diamond p$$

Each logic listed above is complete with respect to its relational semantics. In fact, each of these logics is complete with respect to its finite frames, and therefore has the *finite model property* (see, e.g., Blackburn et al., 2001).

2.2.2 Multi-modal cases. Multi-modal languages are conspicuous in modern applications of modal logic, which often call for combining operators. This happens in a spatial setting, e.g., when describing different topologies at the same time. Such combinations arise by performing certain operations on component logics. Here we recall several basic facts about 'fusion' and 'product' of uni-modal logics. Most of this material can be found in the textbook Gabbay et al., 2003.

The fusion: Let $\mathcal{L}_{\Box_1\Box_2}$ be a bimodal language with modal operators $\Box_1$ and $\Box_2$.

DEFINITION 5.16 *The* fusion *of* **K** *with itself, denoted by* $\mathbf{K} \oplus \mathbf{K}$*, is the least set of formulas of* $\mathcal{L}_{\Box_1\Box_2}$ *containing the axiom* (K) *for both* $\Box_1$ *and* $\Box_2$*, and closed under Modus Ponens,* $\Box_1$*-Necessitation, and* $\Box_2$*-Necessitation.*

The $\mathbf{K} \oplus \mathbf{K}$-frames are triples $\mathfrak{F} = \langle W, R_1, R_2 \rangle$, where W is a nonempty set and R_1 and R_2 are binary relations on W. It is known that $\mathbf{K} \oplus \mathbf{K}$ is complete with respect to this semantics; in fact, it has the finite model property.

We are interested in the fusion of **S4** with itself, which we denote by $\mathbf{S4} \oplus \mathbf{S4}$. It is defined similar to the fusion of **K** with itself. The $\mathbf{S4} \oplus \mathbf{S4}$-frames are triples $\mathfrak{F} = \langle W, R_1, R_2 \rangle$, where W is a nonempty set and R_1 and R_2 are reflexive and transitive. We call such a frame *rooted* if there is a $w \in W$ such that for all $v \in W$ it holds that $w(R_1 \cup R_2)^* v$, where $(R_1 \cup R_2)^*$ is the transitive closure of $R_1 \cup R_2$. It is known that $\mathbf{S4} \oplus \mathbf{S4}$ is complete with respect to this semantics; in fact, $\mathbf{S4} \oplus \mathbf{S4}$ is complete with respect to finite rooted $\mathbf{S4} \oplus \mathbf{S4}$-frames.

Let $\mathcal{T}_{2,2}$ denote the *full infinite quaternary tree* whose each node is R_1-related to two of its four immediate successors and R_2-related to the other two (see Fig. 5.8). We will make use of the next proposition in Sec. 3.2.2.

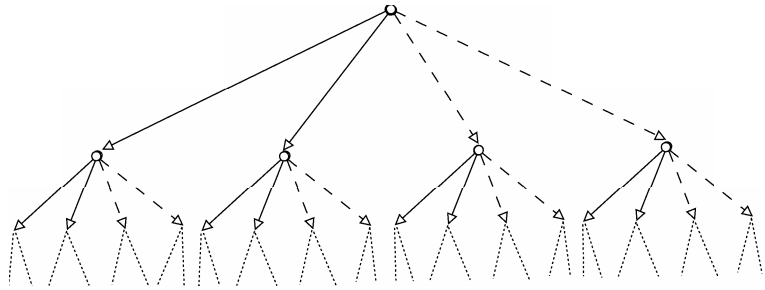


Figure 5.8. $\mathcal{T}_{2,2}$. The solid lines represent R_1 and the dashed lines represent R_2. The dotted lines at the final nodes indicate that the pattern repeats infinitely.

PROPOSITION 5.17 (VAN BENTHEM ET AL., 2005) $\mathbf{S4} \oplus \mathbf{S4}$ *is complete with respect to* $\mathcal{T}_{2,2}$.

The product: For two $\mathbf{K}$-frames $\mathfrak{F} = \langle W, S\rangle$ and $\mathfrak{G} = \langle V, T\rangle$, define the *product frame* $\mathfrak{F} \times \mathfrak{G}$ to be the frame $\langle W \times V, R_1, R_2\rangle$, where for $w, w' \in W$ and $v, v' \in V$:

$$(w, v)R_1(w', v') \text{ iff } wSw' \text{ and } v = v'$$
$$(w, v)R_2(w', v') \text{ iff } w = w' \text{ and } vTv'.$$

The frame $\mathfrak{F} \times \mathfrak{G}$ can be viewed as a $\mathbf{K} \oplus \mathbf{K}$-frame by interpreting the modalities $\Box_1$ and $\Box_2$ of $\mathcal{L}_{\Box_1\Box_2}$ as follows.

$$(w, v) \models \Box_1\varphi \quad \text{iff} \quad \forall(w', v') \text{ if } (w, v)R_1(w', v') \text{ then } (w', v') \models \varphi$$
$$(w, v) \models \Box_2\varphi \quad \text{iff} \quad \forall(w', v') \text{ if } (w, v)R_2(w', v') \text{ then } (w', v') \models \varphi.$$

Let $\mathbf{K} \times \mathbf{K}$ denote the logic of products of $\mathbf{K}$-frames. It is well known that $\mathbf{K} \times \mathbf{K}$ is axiomatized by adding the following two axioms to the fusion $\mathbf{K} \oplus \mathbf{K}$:

$$com = \Box_1\Box_2 p \leftrightarrow \Box_2\Box_1 p$$

$$chr = \Diamond_1\Box_2 p \rightarrow \Box_2\Diamond_1 p.$$

In a similar fashion we define the product of two $\mathbf{S4}$-frames. Let $\mathbf{S4} \times \mathbf{S4}$ denote the logic of products of $\mathbf{S4}$-frames. As with $\mathbf{K} \times \mathbf{K}$, the product logic $\mathbf{S4} \times \mathbf{S4}$ is axiomatized by adding com and chr to the fusion $\mathbf{S4} \oplus \mathbf{S4}$.

2.3 Interpreting $\Box$ as interior and $\Diamond$ as closure

Let $M = \langle \mathcal{X}, \nu\rangle$ be a topo-model, where $\mathcal{X} = \langle X, \tau\rangle$ is a topological space and $\nu : P \rightarrow \mathcal{P}(X)$ is a valuation. We gave the inductive definition of when a formula φ is true at a point x of the model M in Definition 5.1 of Sec. 1.1. The $\Box$ and $\Diamond$ clauses of Definition 5.1 imply that if φ is interpreted as a subset A

of a topological space $\mathcal{X}$, then $\Box\varphi$ stands for $\mathrm{Int}(A)$ and $\Diamond\varphi$ for $\mathrm{Cl}(A)$. Either notion can be used as a primitive. In what follows we emphasize one or the other, depending on the ease of exposition.

DEFINITION 5.18 *We say that φ is* true *in $M = \langle \mathcal{X}, \nu \rangle$ if φ is true at every $x \in X$. We say that φ is* valid *in $\mathcal{X}$ if φ is true in every model based on $\mathcal{X}$. Finally, we say that φ is* valid *in a class of topological spaces if φ is valid in every member of the class.*

EXAMPLE 5.19 Let **Top** denote the class of all topological spaces.

1 First we show that (T) is valid in **Top**. Let $\mathcal{X} \in \mathbf{Top}$, $M = \langle \mathcal{X}, \nu \rangle$ be a topological model, and $x \models \Box p$ for $x \in X$. Then there exists an open neighborhood U of x such that $y \models p$ for each $y \in U$. In particular, since $x \in U$, we obtain that $x \models p$.

2 Next we show that (4) is valid in **Top**. Let $\mathcal{X} \in \mathbf{Top}$, $M = \langle \mathcal{X}, \nu \rangle$ be a topological model, and $x \models \Box p$ for $x \in X$. Then there exists an open neighborhood U of x such that $y \models p$ for each $y \in U$. But then $y \models \Box p$ for each $y \in U$, implying that $x \models \Box\Box p$.

3 Now we show that (K) is valid in **Top**. Let $\mathcal{X} \in \mathbf{Top}$, $M = \langle \mathcal{X}, \nu \rangle$ be a topological model, and for $x \in X$ we have $x \models \Box(p \to q)$ and $x \models \Box p$. Then there exist open neighborhoods U and V of x such that $y \models p \to q$ for each $y \in U$ and $z \models p$ for each $z \in V$. Let $W = U \cap V$. Then W is an open neighborhood of x and for each $w \in W$ we have $w \models p \to q$ and $w \models p$. Therefore, $w \models q$ for each $w \in W$, implying that $x \models \Box q$.

4 Finally, we show that the necessitation rule preserves validity. If $\Box\varphi$ is not valid, then there exists a topological model $M = \langle \mathcal{X}, \nu \rangle$ and $x \in X$ such that $x \not\models \Box\varphi$. Therefore, there exists $y \in X$ such that $y \not\models \varphi$, implying that φ is not valid.

Consequently, we obtain that the modal logic **S4** is sound with respect to interpreting $\Diamond$ as closure. In fact, as shown in McKinsey and Tarski, 1944, **S4** is also complete with respect to this semantics.

2.4 Basic topo-completeness of S4

As we already pointed out, **S4** is sound with respect to interpreting $\Diamond$ as the closure operator of a topological space. We are ready to show that **S4** is in fact complete with respect to this semantics. But first we discuss the well-known connection between relational and topological semantics of **S4** (Aiello et al., 2003; Bezhanishvili and Gehrke, 2005).

2.4.1 Connection with relational semantics of S4.

DEFINITION 5.20 *A topological space $\mathcal{X}$ is called an* Alexandroff *space if the intersection of any family of open subsets of X is again open.*

Equivalently, X is Alexandroff iff every $x \in X$ has a least open neighborhood. There is a close connection between Alexandroff spaces and **S4**-frames. Suppose $\mathfrak{F} = \langle X, R\rangle$ is an **S4**-frame. A subset A of X is called an *upset* of $\mathfrak{F}$ if $x \in A$ and xRy imply $y \in A$. Dually, A is called a *downset* if $x \in A$ and yRx imply $y \in A$.

For a given **S4**-frame $\mathfrak{F} = \langle X, R\rangle$, we define the topology τ_R on X by declaring the upsets of $\mathfrak{F}$ to be open. Then the downsets of $\mathfrak{F}$ turn out to be closed, and it is routine to verify that the obtained space is Alexandroff, that a least neighborhood of $x \in X$ is $R(x) = \{y \in X : xRy\}$, that the closure of a set $A \subseteq X$ is

$$R^{-1}(A) = \{x \in X : \exists y \in A \text{ with } xRy\},$$

and that the interior of A is

$$X - R^{-1}(X - A) = \{x \in X : (\forall y \in X)(xRy \to y \in A)\}.$$

Conversely, for a topological space $\mathcal{X}$ we define the *specialization order* on X by $xR_\tau y$ iff $x \in \mathrm{Cl}(y)$. Then it is routine to check that the specialization order is reflexive and transitive, and that it is a partial order iff $\mathcal{X}$ is T_0. Moreover, one can easily check that $R = R_{\tau_R}$, that $\tau \subseteq \tau_{R_\tau}$, and that $\tau = \tau_{R_\tau}$ iff X is Alexandroff.

These observations immediately imply that there is a 1-1 correspondence between Alexandroff spaces and **S4**-frames, and between Alexandroff T_0-spaces and partially ordered **S4**-frames. Since every finite topological space is an Alexandroff space, this gives a 1-1 correspondence between finite topological spaces and finite **S4**-frames, and between finite T_0-spaces and finite partially ordered **S4**-frames. It is straightforward to see that this also implies a 1-1 correspondence between continuous maps and order-preserving maps, as well as between interior maps and p-morphisms. As a consequence of all this, we obtain the following:

COROLLARY 5.21 *Every normal extension of* **S4** *that is complete with respect to relational semantics is also complete with respect to topological semantics.*

2.4.2 Canonical topo-model of S4. Corollary 5.21 says that standard modal models are a particular case of general topological semantics. Hence, the known completeness of **S4** plus the topological soundness of its axioms immediately give us general topological completeness. Even so, we now give a direct model-theoretic proof of this result, taken from Aiello et al., 2003. It is closely related to the standard modal Henkin construction and is much

like completeness proofs for neighborhood semantics (Chellas, 1980), but with some nice topological twists.

DEFINITION 5.22

1 *Call a set* Γ *of formulas of* $\mathcal{L}$ (**S4**–)consistent *if for no finite set* $\{\varphi_1, \ldots, \varphi_n\} \subseteq \Gamma$ *we have that* $\mathbf{S4} \vdash \neg(\varphi_1 \wedge \cdots \wedge \varphi_n)$.

2 *A consistent set of formulas* Γ *is called* maximally consistent *if there is no consistent set of formulas properly containing* Γ.

It is well known that Γ is maximally consistent iff, for each formula φ of $\mathcal{L}$, either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$, but not both. Now we define a topological space out of maximally consistent sets of formulas.

DEFINITION 5.23 (CANONICAL TOPOLOGICAL SPACE) *The* canonical topological space *is the pair* $\mathcal{X}^{\mathcal{L}} = \langle X^{\mathcal{L}}, \tau^{\mathcal{L}} \rangle$ *where:*

- $X^{\mathcal{L}}$ *is the set of all maximally consistent sets;*
- $\tau^{\mathcal{L}}$ *is the set generated by arbitrary unions of the following* basic sets $B^{\mathcal{L}} = \{\widehat{\Box\varphi} : \varphi \text{ is any formula}\}$, *where* $\widehat{\varphi} =_{def} \{x \in X^{\mathcal{L}} : \varphi \in x\}$. *In other words, basic sets are the families of the form:* $U_\varphi = \{x \in X^{\mathcal{L}} : \Box\varphi \in x\}$.

We first check that $\mathcal{X}^{\mathcal{L}}$ is indeed a topological space.

LEMMA 5.24 $B^{\mathcal{L}}$ *forms a basis for the topology.*

Proof We only need to show the following two properties:

- For each $U_\varphi, U_\psi \in B^{\mathcal{L}}$ and each $x \in U_\varphi \cap U_\psi$, there is $U_\chi \in B^{\mathcal{L}}$ such that $x \in U_\chi \subseteq U_\varphi \cap U_\psi$;
- For each $x \in X^{\mathcal{L}}$, there is $U_\varphi \in B^{\mathcal{L}}$ such that $x \in U_\varphi$.

The necessitation rule implies that $\Box\top \in x$ for each x. Hence, $X^{\mathcal{L}} = \widehat{\Box\top}$, and so the second item is satisfied. As for the first item, thanks to the axiom (K), one can easily check that $\widehat{\Box(\varphi \wedge \psi)} = \widehat{\Box\varphi} \cap \widehat{\Box\psi}$. Hence, $U_\varphi \cap U_\psi \in B^{\mathcal{L}}$, and so $B^{\mathcal{L}}$ is closed under finite intersections; whence, the first property is satisfied. QED

Next we define the canonical topo-model.

DEFINITION 5.25 (CANONICAL TOPO-MODEL) *The* canonical topo-model *is the pair* $M^{\mathcal{L}} = \langle \mathcal{X}^{\mathcal{L}}, \nu^{\mathcal{L}} \rangle$ *where:*

- $\mathcal{X}^{\mathcal{L}}$ *is the canonical topological space;*

- $\nu^{\mathcal{L}}(p) = \{x \in X^{\mathcal{L}} : p \in x\}$.

The valuation $\nu^{\mathcal{L}}$ equates truth of a propositional letter *at* a maximally consistent set with its membership *in* that set. We now show this harmony between the two viewpoints lifts to all formulas.

LEMMA 5.26 (TRUTH LEMMA) For all modal formulas φ,

$$M^{\mathcal{L}}, x \models_{\mathcal{L}} \varphi \text{ iff } x \in \widehat{\varphi}.$$

Proof Induction on the complexity of φ. The base case follows from the definition above. The case of the booleans follows from the following well-known identities for maximally consistent sets:

- $\widehat{\neg\varphi} = X^{\mathcal{L}} - \widehat{\varphi}$;
- $\widehat{\varphi \wedge \psi} = \widehat{\varphi} \cap \widehat{\psi}$.

The interesting case is that of the modal operator $\Box$. We do the two relevant implications separately, starting with the easy one.

$\Leftarrow$ 'From membership to truth.' Suppose $x \in \widehat{\Box\varphi}$. By definition, $\widehat{\Box\varphi}$ is a basic set, hence open. Moreover, thanks to the axiom (T), we have $\widehat{\Box\varphi} \subseteq \widehat{\varphi}$. Therefore, there exists an open neighborhood $U = \widehat{\Box\varphi}$ of x such that $y \in \widehat{\varphi}$, for any $y \in U$, and by the induction hypothesis, $M^{\mathcal{L}}, y \models_{\mathcal{L}} \varphi$. Thus $M^{\mathcal{L}}$, $x \models_{\mathcal{L}} \Box\varphi$.

$\Rightarrow$ 'From truth to membership.' Suppose $M^{\mathcal{L}}, x \models_{\mathcal{L}} \Box\varphi$. Then there exists a basic set $\widehat{\Box\psi} \in B^{\mathcal{L}}$ such that $x \in \widehat{\Box\psi}$ and $M^{\mathcal{L}}, y \models_{\mathcal{L}} \varphi$ for all $y \in \widehat{\Box\psi}$. By the induction hypothesis, $y \in \widehat{\varphi}$ for all $y \in \widehat{\Box\psi}$. Therefore, $\widehat{\Box\psi} \subseteq \widehat{\varphi}$. This implies that **S4** can prove the implication $\Box\psi \to \varphi$. But then **S4** can prove $\Box\Box\psi \to \Box\phi$, and hence, using the axiom (4), it can also prove $\Box\psi \to \Box\phi$. It follows that $\widehat{\Box\psi} \subseteq \widehat{\Box\phi}$, and so the world x belongs to $\widehat{\Box\phi}$. QED

Now we can clinch the proof of our main result.

THEOREM 5.27 (COMPLETENESS) *For any set of formulas* Γ,

$$\textit{if} \quad \Gamma \models_{\mathcal{L}} \varphi \quad \textit{then} \quad \Gamma \vdash_{\mathbf{S4}} \varphi.$$

Proof Suppose that $\Gamma \not\vdash_{\mathbf{S4}} \varphi$. Then $\Gamma \cup \{\neg\varphi\}$ is consistent, and by the standard Lindenbaum lemma argument it can be extended to a maximally consistent set x. By the truth lemma, $M^{\mathcal{L}}, x \models_{\mathcal{L}} \neg\varphi$, whence $M^{\mathcal{L}}, x \not\models_{\mathcal{L}} \varphi$, and we have constructed the required counter-model. QED

COROLLARY 5.28 **S4** *is the logic of the class of all topological spaces.*

We note that the whole construction in the completeness proof above would also work if we restricted attention to the *finite* language consisting of the initial formula and all its subformulas. This means that we only get finitely many maximally consistent sets, and so non-provable formulas can be refuted on *finite models*, whose size is effectively computable from the formula itself. This is usually referred to as the *effective finite model property*.

COROLLARY 5.29

1 **S4** *is the logic of the class of all* finite *topological spaces.*

2 **S4** *has the* effective finite model property *with respect to the class of topological spaces.*

Incidentally, this also shows that validity in **S4** is *decidable*, but we forego such complexity issues in this chapter.

Comparing our construction with the standard modal Henkin model $\langle X^{\mathcal{L}}, R^{\mathcal{L}}, \models_{\mathcal{L}} \rangle$ for **S4**, the basic sets of our topology $\tau^{\mathcal{L}}$ are $R^{\mathcal{L}}$-upward closed. Hence, every open of $\mathcal{X}^{\mathcal{L}}$ is $R^{\mathcal{L}}$-upward closed, and $\tau^{\mathcal{L}}$ is weaker than the topology $\tau_{R^{\mathcal{L}}}$ corresponding to $R^{\mathcal{L}}$. In particular, our canonical topological space is *not* an Alexandroff space. In fact, as shown in Gabelaia, 2001, Theorem 3.2.3, $\tau_{R^{\mathcal{L}}}$ is the least Alexandroff topology containing $\tau_{\mathcal{L}}$. Further discussion of this and related topics can be found in Aiello et al., 2003, Sec. 3 and Gabelaia, 2001.

2.5 Completeness in special spaces

We have already seen that **S4** is the logic of all topological spaces. But there are classical results with much more mathematical content such as McKinsey and Tarski's beautiful theorem that **S4** is also the logic of any dense-in-itself metric separable space. Here we concentrate on three spaces that play an important role in mathematics—the Cantor space $\mathbb{C}$, the rational line $\mathbb{Q}$, and the real line $\mathbb{R}$—and sketch three proofs, taken respectively from Aiello et al., 2003, van Benthem et al., 2005 and Bezhanishvili and Gehrke, 2005, that **S4** is the logic of each of these spaces.

2.5.1 Completeness w.r.t. $\mathbb{C}$. We first show that **S4** is the logic of the Cantor space $\mathbb{C}$. Our exposition is rather sketchy. For full details we refer the reader to Aiello et al., 2003, Sec. 4.1.

Suppose an **S4**-frame $\mathfrak{F} = \langle W, R \rangle$ is given. We recall that $\mathfrak{F}$ is *rooted* if there exists $r \in W$—called a *root* of $\mathfrak{F}$—such that rRw for each $w \in W$. We call $C \subseteq W$ a *cluster* if for all $w, v \in C$ we have wRv and vRw. A cluster C is called *simple* if it consists of a single point, and *proper* if it consists of more than one point. The next theorem will aid in proving that **S4** is the logic of $\mathbb{C}$.

THEOREM 5.30 (AIELLO ET AL., 2003) **S4** *is complete with respect to finite rooted* **S4***-frames whose every cluster is proper.*

Now let a formula φ be not provable in **S4**. By Theorem 5.30, φ can be refuted in a finite rooted **S4**-model $M = \langle W, R, \nu \rangle$, with a root r, whose every cluster is proper. We transform the latter into a counterexample on the Cantor space $\mathscr{C}$. Our technique is *selective unravelling*, a refinement of the technique of *unravelling* in modal logic (see, e.g., Blackburn et al., 2001). We select those infinite paths of M that are in a 1-1 correspondence with infinite paths of the full infinite binary tree $\mathcal{T}_2$ (see Fig. 5.9).

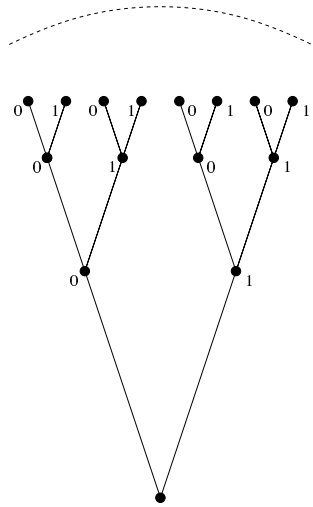


Figure 5.9. $\mathcal{T}_2$.

We start with a root r and announce (r) as a selective path. Then if $(w_1, \ldots, w_k)$ is already a selective path, we introduce a *left* move by announcing $(w_1, \ldots, w_k, w_k)$ as a selective path; and we introduce a *right* move by announcing $(w_1, \ldots, w_k, w_{k+1})$ as a selective path if $w_k R w_{k+1}$ and $w_k \neq w_{k+1}$. (Since we assumed that every cluster of W is proper, such w_{k+1} exists for every w_k.) We call an infinite path σ of W *selective* if every initial segment of σ is a finite selective path of W. We denote by Σ the set of all infinite selective paths of W. For a finite selective path $(w_1, \ldots, w_k)$, let

$$B_{(w_1,\ldots,w_k)} = \{\sigma \in \Sigma : \ \sigma \text{ has an initial segment } (w_1, \ldots, w_k)\}.$$

Define a topology τ_Σ on Σ by introducing

$$\mathcal{B}_\Sigma = \{B_{(w_1,\ldots,w_k)} : (w_1, \ldots, w_k) \text{ is a finite selective path of } W\}$$

as a basis.

To see that $\mathcal{B}_\Sigma$ is a basis, observe that $B_{(r)} = \Sigma$, and that

$$B_{(w_1,\dots,w_k)} \cap B_{(v_1,\dots,v_m)} = \begin{cases} B_{(w_1,\dots,w_k)} & \text{if } (v_1,\dots,v_m) \text{ is an initial segment of } (w_1,\dots,w_k), \\ B_{(v_1,\dots,v_m)} & \text{if } (w_1,\dots,w_k) \text{ is an initial segment of } (v_1,\dots,v_m), \\ \emptyset & \text{otherwise.} \end{cases}$$

In order to define ν_Σ, note that every infinite selective path σ of W either gets stable or keeps cycling. In other words, either $\sigma = (w_1, \dots, w_k, w_k, \dots)$ or $\sigma = (w_1, \dots, w_n, w_{n+1}, \dots)$ where w_i belongs to some cluster $C \subseteq W$ for $i > n$. In the former case we say that w_k *stabilizes* σ, and in the latter that σ *keeps cycling* in C. Now define ν_Σ on Σ by

$$\sigma \in \nu_\Sigma(p) \text{ iff } \begin{cases} w_k \models p & \text{if } w_k \text{ stabilizes } \sigma, \\ \rho(C) \models p & \text{if } \sigma \text{ keeps cycling in } C \subseteq W, \text{ where } \rho(C) \text{ is some arbitrarily chosen representative of } C. \end{cases}$$

All we need to show is that $\langle \Sigma, \tau_\Sigma \rangle$ is homeomorphic to the Cantor space, and that $M_\Sigma = \langle \Sigma, \tau_\Sigma, \nu_\Sigma \rangle$ is topo-bisimilar to the initial M. In order to show the first claim, let us recall that the Cantor space is homeomorphic to the countable topological product of the two element set $\mathbf{2} = \{0, 1\}$ with the discrete topology. To picture the Cantor space, one can think of the full infinite binary tree $\mathcal{T}_2$; starting at the root, one associates 0 to every left-son of a node and 1 to every right-son. Then points of the Cantor space are infinite paths of $\mathcal{T}_2$. This together with the construction of Σ immediately gives us that $\langle \Sigma, \tau_\Sigma \rangle$ is homeomorphic to $\mathcal{C}$.

Finally, we show that M_Σ is topo-bisimilar to M. Define $F : \Sigma \to W$ by

$$F(\sigma) = \begin{cases} w_k & \text{if } w_k \text{ stabilizes } \sigma, \\ \rho(C) & \text{if } \sigma \text{ keeps cycling in } C. \end{cases}$$

Obviously F is well-defined, and is actually surjective. (For any $w_k \in W$, we have $F(\sigma_0, w_k, w_k, \dots) = w_k$, where σ_0 is a selective path from w_1 to w_k.)

PROPOSITION 5.31 *F is a total topo-bisimulation between $M_\Sigma = \langle \Sigma, \tau_\Sigma, \nu_\Sigma \rangle$ and $M = \langle W, R, \nu \rangle$.*

Proof (Sketch) With $\langle W, R \rangle$ we can associate the finite topological space $\langle W, \tau_R \rangle$. The set $\{R(v) : v \in W\}$ forms a basis for τ_R. Now the function $F : \langle \Sigma, \tau_\Sigma \rangle \to \langle W, \tau_R \rangle$ is continuous because

$$F^{-1}(R(v)) = \bigcup \{B_{(w_1,\dots,w_k)} : vRw_k\}$$

for each $v \in W$, and F is open because

$$F(B_{(w_1,\ldots,w_k)}) = R(w_k)$$

for each basic open $B_{(w_1,\ldots,w_k)}$ of $\langle \Sigma, \tau_\Sigma \rangle$. Therefore, F is an interior map. Moreover, as follows from the definition of ν_Σ,

$$\sigma \in \nu_\Sigma(p) \text{ iff } F(\sigma) \in \nu(p).$$

Since every interior map satisfying this condition is a topo-bisimulation, so is our F. QED

THEOREM 5.32 **S4** *is the logic of* $\mathbb{C}$.

Proof Suppose **S4**$\not\vdash \varphi$. Then there is a finite rooted model M such that every cluster of M is proper and M refutes φ. Since $\mathbb{C}$ is homeomorphic to $\langle \Sigma, \tau_\Sigma \rangle$, by Proposition 5.31, there exists a valuation $\nu_{\mathbb{C}}$ on $\mathbb{C}$ such that $\langle \mathbb{C}, \nu_{\mathbb{C}} \rangle$ is topo-bisimilar to M. Hence, φ is refuted on $\mathbb{C}$. QED

2.5.2 Completeness w.r.t. $\mathbb{Q}$. Now we show that **S4** is also the logic of the rational line $\mathbb{Q}$. Our proof is taken from van Benthem et al., 2005. Again we use the infinite binary tree $\mathcal{T}_2$, but this time we work with its nodes rather than infinite paths. We start by recalling the following two well-known results.

THEOREM 5.33 (VAN BENTHEM-GABBAY) **S4** *is complete with respect to* $\mathcal{T}_2$.

Proof For a proof see, e.g., Goldblatt, 1980, Theorem 1 and the subsequent discussion. QED

THEOREM 5.34 (CANTOR) *Every countable dense linear ordering without endpoints is isomorphic to* $\mathbb{Q}$.

Proof For a proof see, e.g., Kuratowski and Mostowski, 1976, p. 217, Theorem 2. QED

REMARK 5.35 We recall that if $\langle X, < \rangle$ is a linearly ordered set and $x, y \in X$ with $x < y$, then the *open interval* (x, y) is the set $\{z \in X \,:\, x < z < y\}$. If we view linearly ordered sets as topological spaces using the set of open intervals as a basis for the topology, then it follows from Cantor's theorem that every countable dense linear ordering without endpoints is (as a topological space) homeomorphic to $\mathbb{Q}$.

We are now ready to proceed with the proof.

THEOREM 5.36 **S4** *is complete with respect to* $\mathcal{CQ}$.

Proof Our strategy is as follows. We use completeness of **S4** with respect to $\mathcal{T}_2$, view $\mathcal{T}_2$ as an Alexandroff space, define a dense subset X of $\mathcal{CQ}$ without endpoints, and establish a topo-bisimulation between X and $\mathcal{T}_2$. This will allow us to transfer counterexamples from $\mathcal{T}_2$ to X, which by Cantor's theorem is order-isomorphic, and hence homeomorphic to $\mathcal{CQ}$.

Let $X = \bigcup_{n\in\omega} X_n$, where $X_0 = \{0\}$ and

$$X_{n+1} = X_n \cup \{x - \frac{1}{3^n}, x + \frac{1}{3^n} \ : \ x \in X_n\}$$

CLAIM 5.37 *For* $n > 0$ *and* $x, y \in X_n$, $x \neq y$ *implies* $|x - y| \geq \frac{1}{3^{n-1}}$.

Proof By induction on n. If $n = 1$, then $X_1 = \{0, 1, -1\}$, and so $x \neq y$ implies $|x - y| \geq 1$. That the claim holds for $n = k + 1$ is also not hard to see. Note that if $u, v \in X_{n-1}$ with $u \neq v$, then, by the induction hypothesis, $|u - v| \geq \frac{1}{3^{n-2}}$ and hence $|(u + \frac{1}{3^{n-1}}) - (v - \frac{1}{3^{n-1}})| \geq \frac{1}{3^{n-1}}$. QED

It follows from Claim 5.37 that $\langle X, <\rangle$ is a countable dense linear ordering without endpoints, thus order-isomorphic, and hence homeomorphic to $\mathcal{CQ}$. It also follows that for each $x \in X$ with $x \neq 0$ there exists n_x with $x \in X_{n_x}$ and $x \notin X_{n_x-1}$, and that there is a unique $y \in X_{n_x-1}$ with $x = y - \frac{1}{3^{n_x-1}}$ or $x = y + \frac{1}{3^{n_x-1}}$. Therefore, the open X-intervals $(x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}})$ form a basis for the order-topology on X.

Now we define f from X onto $\mathcal{T}_2$ by recursion (see Fig. 5.10): If $x = 0$ then we let $f(0)$ be the root r of $\mathcal{T}_2$; if $x \neq 0$ then $x \in X_{n_x} - X_{n_x-1}$ and we let

$$f(x) = \begin{cases} \text{the immediate left successor of } f(y) & \text{if } x = y - \frac{1}{3^{n_x-1}} \\ \text{the immediate right successor of } f(y) & \text{if } x = y + \frac{1}{3^{n_x-1}} \end{cases} .$$

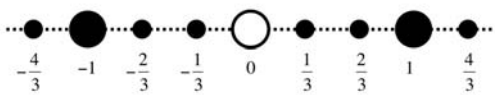


Figure 5.10. The first stages of the labelling in the completeness proof for **S4**: 0 is labelled by the root r of $\mathcal{T}_2$, -1 is labelled by the immediate left successor of r, 1 is labelled by the immediate right successor of r, and so on.

CLAIM 5.38 f *is an interior map.*

Proof (Sketch) We recall that a basis for the Alexandroff topology on $\mathcal{T}_2$ is $\mathcal{B} = \{B_t\}_{t \in \mathcal{T}_2}$ where $B_t = \{s \in \mathcal{T}_2 : tRs\}$. Now f is open because for a basic open X-interval $(x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}})$, we have $f(x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}}) = B_{f(x)}$. Also f is continuous because for each $t \in \mathcal{T}_2$, the f-inverse image of B_t is open. Indeed, if $x \in f^{-1}(B_t)$, then $f(x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}}) = B_{f(x)} \subseteq B_t$, implying that there exists an open interval $I = (x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}})$ of x such that $I \subseteq f^{-1}(B_t)$. Thus, f is interior. QED

To complete the proof, if **S4** $\not\vdash \varphi$, then by Theorem 5.33, there is a valuation ν on $\mathcal{T}_2$ such that $\langle \mathcal{T}_2, \nu \rangle, r \not\models \varphi$. Define a valuation ξ on X by $\xi(p) = f^{-1}(\nu(p))$. Since f is an interior map, and $f(0) = r$, we have that 0 and r are topo-bisimilar. Therefore, $\langle X, \xi \rangle, 0 \not\models \varphi$. Now since X is homeomorphic to $\mathbb{Q}$, we obtain that φ is also refutable on $\mathbb{Q}$. QED

2.5.3 Completeness w.r.t. $\mathbb{R}$. Finally, we show that **S4** is also the logic of the real line $\mathbb{R}$. There are at least three different proofs of this result. The original one is a particular case of a more general theorem (McKinsey and Tarski, 1944) that **S4** is the logic of any dense-in-itself metric separable space (see also Rasiowa and Sikorski, 1963). The other two can be found in Aiello et al., 2003 and Bezhanishvili and Gehrke, 2005. Here we sketch the proof given in Bezhanishvili and Gehrke, 2005, where the construction of the Cantor set on any bounded interval of $\mathbb{R}$ is used to show that every finite rooted **S4**-frame is an interior image of $\mathbb{R}$.

Suppose $a, b \in \mathbb{R}$, $a < b$, and $I = (a, b)$. We recall that the Cantor set $\mathbb{C}$ is constructed inside I by taking out open intervals from I infinitely many times. More precisely, in step 1 of the construction the open interval

$$I_1^1 = (a + \frac{b-a}{3}, a + \frac{2(b-a)}{3})$$

is taken out. We denote the remaining closed intervals by J_1^1 and J_2^1. In step 2 the open intervals

$$I_1^2 = (a + \frac{b-a}{3^2}, a + \frac{2(b-a)}{3^2}) \text{ and } I_2^2 = (a + \frac{7(b-a)}{3^2}, a + \frac{8(b-a)}{3^2})$$

are taken out. We denote the remaining closed intervals by J_1^2, J_2^2, J_3^2, and J_4^2. In general, in step m the open intervals $I_1^m, \ldots, I_{2^{m-1}}^m$ are taken out, and the closed intervals $J_1^m, \ldots, J_{2^m}^m$ remain.

Our immediate goal is to show that every finite tree is an interior image of I. We first show that the tree T of depth 2 and branching n shown in Fig. 5.11 is an interior image of I, and then extend this result to any finite tree by induction on the depth of the tree.

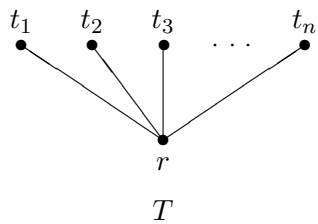


Figure 5.11. The tree of depth 2 and branching n.

LEMMA 5.39 *T is an interior image of I.*

Proof Define $f_I^T : I \to T$ by

$$f_I^T(x) = \begin{cases} t_k, & \text{if } x \in \bigcup_{m \equiv k (mod\ n)} \bigcup_{p=1}^{2^{m-1}} I_p^m \\ r, & \text{otherwise.} \end{cases}$$

Obviously, f_I^T is a well-defined onto map. Moreover,

$$(f_I^T)^{-1}(t_k) = \bigcup_{m \equiv k (mod\ n)} \bigcup_{p=1}^{2^{m-1}} I_p^m \quad \text{and} \quad (f_I^T)^{-1}(r) = \mathcal{C}.$$

Since $\{\emptyset, \{t_1\}, \ldots, \{t_n\}, T\}$ is a family of basic open subsets of T, it is obvious that f_I^T is continuous. Suppose U is an open interval of I. If $U \cap \mathcal{C} = \emptyset$, then $f_I^T(U) \subseteq \{t_1, \ldots, t_n\}$, and so $f_I^T(U)$ is open. If $U \cap \mathcal{C} \neq \emptyset$, then there exists $c \in U \cap \mathcal{C}$. Since $c \in \mathcal{C}$, we have $f_I^T(c) = r$. From $c \in U$ it follows that there is $\varepsilon > 0$ such that $(c - \varepsilon, c + \varepsilon) \subseteq U$. We pick m so that $\frac{b-a}{3^m} < \varepsilon$. As $c \in \mathcal{C}$, there is $k \in \{1, \ldots, 2^m\}$ such that $c \in J_k^m$. Moreover, since the length of J_k^m is equal to $\frac{b-a}{3^m}$, we have that $J_k^m \subseteq U$. Therefore, U contains the points removed from J_k^m in the subsequent iterations in the construction of $\mathcal{C}$. Thus, $f_I^T(U) \supseteq \{t_1, \ldots, t_n\}$ and $f_I^T(U) = T$. Hence, $f_I^T(U)$ is open for any open interval U of I. It follows that f_I^T is an onto interior map. QED

THEOREM 5.40 *Every finite tree of branching $n \geq 1$ is an interior image of I.*

Proof Suppose T is a finite tree of branching $n \geq 1$. Without loss of generality we may assume that the depth of T is $d+1$, where $d \geq 2$. Then we can represent T as shown in Fig. 5.12, where $t_1, \ldots, t_{n^d}$ are the elements of T of depth 2, and T_d is the subtree of T of all elements of T of depth ≥ 2. We note that for each $k \in \{1, \ldots, n^d\}$ the upset $R(t_k)$ is isomorphic to the tree of depth 2

and branching n, and that T_d is the tree of depth d and branching n. So by the induction hypothesis, there is an onto interior map from I onto T_d. Also, by Lemma 5.39, there exists an onto interior map from I onto each $R(t_k)$. Now putting these maps together produces an onto interior map from I onto T. For the details we refer to Bezhanishvili and Gehrke, 2005, Theorem 8. QED

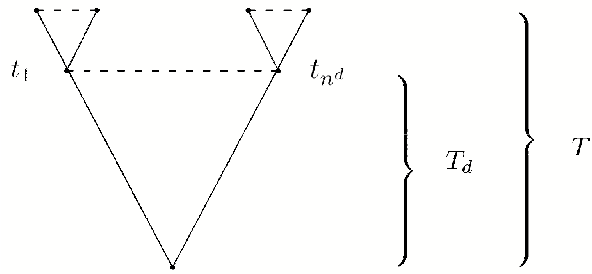


Figure 5.12. T and T_d.

The following useful assertion is now an easy consequence.

COROLLARY 5.41 *Every finite rooted partially ordered* **S4**-*frame is an interior image of* $\mathbb{R}$.

Proof Since every finite rooted partially ordered **S4**-frame is a p-morphic image of some finite tree of branching $n \geq 1$ (Bezhanishvili and Gehrke, 2005, Lemma 4), it follows from Theorem 5.40 that every finite rooted partially ordered **S4**-frame is an interior image of any bounded open interval $I \subseteq \mathbb{R}$. Since I is homeomorphic to $\mathbb{R}$, the corollary follows. QED

We now extend on Corollary 5.41 and show that all finite rooted **S4**-frames are interior images of $\mathbb{R}$. Suppose $\mathfrak{F}$ is an **S4**-frame. We define an equivalence relation $\sim$ on $\mathfrak{F}$ by $w \sim v$ iff w, v belong to the same cluster. Let $\mathfrak{F}/\!\sim$ denote the *skeleton* of $\mathfrak{F}$. That is, $\mathfrak{F}/\!\sim$ is the quotient of $\mathfrak{F}$ by $\sim$, and $R_\sim$ is defined on $W/\!\sim$ componentwise.

DEFINITION 5.42 *We call an* **S4**-*frame* $\mathfrak{F}$ *a* quasi-tree *if* $\mathfrak{F}/\!\sim$ *is a tree.*

Suppose Q is a quasi-tree. We say that the *swelling* of Q is q if every cluster of Q consists of exactly q elements. Again, our immediate goal is to show that every finite quasi-tree is an interior image of I. This we show by first obtaining the quasi-tree Q of depth 2, branching n, and swelling q, shown in Fig. 5.13, as an interior image of I, and then extending this result to any finite quasi-tree by induction on the depth of the quasi-tree. For this we use the following lemma taken from Bezhanishvili and Gehrke, 2005, Lemma 11.

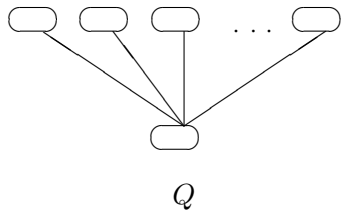


Figure 5.13. Quasi-tree of depth 2, branching n, and swelling q.

LEMMA 5.43 *If X has a countable basis and every countable subset of X is boundary, then for each natural number n there exist disjoint dense and boundary subsets $A_1, \ldots, A_n$ of X such that $X = \bigcup_{i=1}^{n} A_i$.*

LEMMA 5.44 *Q is an interior image of I.*

Proof We denote the least cluster of Q by r and its elements by $r_1, \ldots, r_q$. Also for $1 \leq i \leq n$ we denote the ith maximal cluster of Q by t^i and its elements by $t^i_1, \ldots, t^i_q$. Since the Cantor set $\mathbb{C}$ satisfies the conditions of Lemma 5.43, it can be divided into q-many disjoint dense and boundary subsets $\mathbb{C}_1, \ldots, \mathbb{C}_q$. Also each I^m_p ($1 \leq p \leq 2^{m-1}$, $m \in \omega$) satisfies the conditions of Lemma 5.43, and so each I^m_p can be divided into q-many disjoint dense and boundary subsets $(I^m_p)^1, \ldots, (I^m_p)^q$. Suppose $1 \leq k \leq q$. We define $f^Q_I : I \to Q$ by

$$f^Q_I(x) = \begin{cases} t^i_k, & \text{if } x \in \bigcup_{m \equiv i (mod\ n)} \bigcup_{p=1}^{2^{m-1}} (I^m_p)^k \\ r_k, & \text{if } x \in \mathbb{C}_k. \end{cases}$$

It is clear that f^Q_I is a well-defined onto map. As with Lemma 5.39, we have

$$(f^Q_I)^{-1}(t^i) = \bigcup_{m \equiv i (mod\ n)} \bigcup_{p=1}^{2^{m-1}} I^m_p \quad \text{and} \quad (f^Q_I)^{-1}(r) = \mathbb{C}.$$

Hence, f^Q_I is continuous. To show that f^Q_I is open let U be an open interval in I. If $U \cap \mathbb{C} = \emptyset$, then $f^Q_I(U) \subseteq \bigcup_{i=1}^{n} t^i$. Moreover, since $(I^m_p)^1, \ldots, (I^m_p)^q$ partition I^m_p into q-many disjoint dense and boundary subsets, $U \cap I^m_p \neq \emptyset$ implies $U \cap (I^m_p)^k \neq \emptyset$ for every $k \in \{1, \ldots, q\}$. Hence, if $f^Q_I(U)$ contains an element of a cluster t^i, it contains the whole cluster. Thus, $f^Q_I(U)$ is open. Now suppose $U \cap \mathbb{C} \neq \emptyset$. Since $\mathbb{C}_1, \ldots, \mathbb{C}_q$ partition $\mathbb{C}$ into q-many disjoint dense and boundary subsets, $U \cap \mathbb{C}_k \neq \emptyset$ for every $k \in \{1, \ldots, q\}$. Hence, $r \subseteq f^Q_I(U)$. Moreover, the same argument as in the proof of Lemma 5.39

guarantees that every point greater than points in r also belongs to $f_I^Q(U)$. Thus $f_I^Q(U) = Q$, implying that f_I^Q is an onto interior map. QED

THEOREM 5.45 *Every finite quasi-tree of branching n and swelling q is an interior image of I.*

Proof This follows along the same lines as the proof of Theorem 5.40 but is based on Lemma 5.44 instead of Lemma 5.39. QED

COROLLARY 5.46 *Every finite rooted* **S4***-frame is an interior image of $\mathbb{R}$.*

Proof This follows along the same lines as the proof of Corollary 5.41 but is based on the fact that every finite rooted **S4**-frame is a p-morphic image of some finite quasi-tree of branching n and swelling q (Bezhanishvili and Gehrke, 2005, Lemma 5) and Theorem 5.45. QED

THEOREM 5.47 **S4** *is complete with respect to $\mathbb{R}$.*

Proof If **S4** $\nvdash \varphi$, then there exists a finite rooted **S4**-model $M = \langle W, R, \nu \rangle$, with a root r, such that $M, r \not\models \varphi$. By Corollary 5.46, there exists an onto interior map $f : \mathbb{R} \to W$. Define a valuation ξ on $\mathbb{R}$ by $\xi(p) = f^{-1}(\nu(p))$. Then f is a total topo-bisimulation between $\langle \mathbb{R}, \xi \rangle$ and M. Thus, there exists a point $x \in \mathbb{R}$ such that $x \not\models \varphi$. QED

We recall that a subset A of $\mathbb{R}$ is *convex* if $x, y \in A$ and $x \leq z \leq y$ imply that $z \in A$.

COROLLARY 5.48 **S4** *is complete with respect to boolean combinations of countable unions of convex subsets of $\mathbb{R}$.*

Proof Let $f : \mathbb{R} \to W$ be the onto interior map from the proof of Theorem 5.47. Observe that for each $w \in W$ we have that $f^{-1}(w)$ is a boolean combination of countable unions of convex subsets of $\mathbb{R}$ (Bezhanishvili and Gehrke, 2005, Theorem 15). The result follows. QED

2.6 The landscape of spatial logics over S4

As we have already seen, **S4** is the logic of all topological spaces when interpreting $\Diamond$ as closure. In addition, **S4** turned out to be the logic of the Cantor space $\mathbb{C}$, the rational line $\mathbb{Q}$, the real line $\mathbb{R}$, or more generally, any dense-in-itself metric separable space. These results, although with a lot of mathematical content, also indicate serious limitations of the basic modal language

in expressing various topological properties. For example, the completeness of **S4** with respect to any dense-in-itself metric separable space already implies that such topological properties as being dense-in-itself, metric, or separable are not definable in the basic modal language. Here we address the topological definability issue, as well as review several normal extensions of **S4** that are complete with respect to interesting classes of topological spaces.

Topological definability and undefinability: Suppose a class K of topological spaces is given. We say that K is *topologically definable* or simply *topo-definable* if there exists a set of modal formulas Γ such that for each topological space $\mathcal{X}$ we have $\mathcal{X} \in K$ iff $\mathcal{X} \models \Gamma$. Topological completeness of **S4** tells us that the class **Top** of *all* topological spaces is topo-definable (by the formula $\top$ over **S4**). However, as we will see below, many important classes of topological spaces such as the classes of compact or connected spaces are *not* topo-definable. For this we will need the following theorem, first established in Gabelaia, 2001 and van Benthem et al., 2003.

THEOREM 5.49 *Suppose φ is an arbitrary modal formula.*

1. *If $\mathcal{Y}$ is an interior image of $\mathcal{X}$, then $\mathcal{X} \models \varphi$ implies $\mathcal{Y} \models \varphi$.*
2. *If $\mathcal{Y}$ is an open subspace of $\mathcal{X}$, then $\mathcal{X} \models \varphi$ implies $\mathcal{Y} \models \varphi$.*
3. *If $\mathcal{X}$ is the topological sum of $\{\mathcal{X}_i\}_{i \in I}$, then $\mathcal{X} \models \varphi$ iff $\mathcal{X}_i \models \varphi$ for each $i \in I$.*

Now we are ready to show that compactness, connectedness, and the separation axioms T_0, T_d, T_1, and T_2 are not topo-definable.

PROPOSITION 5.50 (GABELAIA, 2001)

1. *Neither compactness nor connectedness is topo-definable.*
2. *None of the separation axioms T_0, T_d, T_1, and T_2 is topo-definable.*

Proof (1) Let $\mathcal{X} = \langle \{x\}, \tau \rangle$ be a singleton set with the discrete topology. Then obviously $\mathcal{X}$ is both compact and connected. On the other hand, any infinite topological sum of $\mathcal{X}$ is neither compact nor connected. Now apply Theorem 5.49(3).

(2) Let $\mathcal{X} = \langle \{x, y\}, \tau \rangle$ be a two point set with the trivial topology. Then obviously $\mathcal{X}$ does not satisfy any of the four separation axioms. Define $f : \mathbb{R} \to \{x, y\}$ by

$$f(r) = \begin{cases} x, & \text{if } r \in \mathbb{Q} \\ y, & \text{otherwise.} \end{cases}$$

Then it is easy to see that f is an onto interior map. Now observe that $\mathbb{R}$ satisfies all the four separation axioms and apply Theorem 5.49(1). QED

REMARK 5.51 $\mathbb{R}$ also satisfies stronger separation axioms such as T_3 (regularity), $T_{3\frac{1}{2}}$ (complete regularity), T_4 (normality), T_5, and T_6. Therefore, Proposition 5.50 also implies that none of $T_3, T_{3\frac{1}{2}}, T_4, T_5$, and T_6 is topo-definable.

Proposition 5.50 indicates that the basic modal language $\mathcal{L}$ is not expressive enough for topological purposes. In Sec. 3 we will consider several enrichments of $\mathcal{L}$ and show that some of topological properties not expressible in $\mathcal{L}$ can be expressed in its various enrichments. Nevertheless, it seems to be a natural question to characterize those classes of topological spaces that *can* be defined in $\mathcal{L}$. An answer to this question was given in Gabelaia, 2001, where a topological analogue of the Goldblatt-Thomason theorem was established.

DEFINITION 5.52 (GABELAIA, 2001) *Let $\mathcal{X} = \langle X, \tau \rangle$ be a topological space. We let* $\mathrm{uf}(X)$ *denote the set of ultrafilters of the powerset $\mathcal{P}(X)$ and define R on* $\mathrm{uf}(X)$ *by*

$$wRu \text{ iff } A \in u \text{ implies } \mathrm{Cl}(A) \in w$$

for each $A \subseteq X$. It is easy to verify that R is reflexive and transitive on $\mathrm{uf}(X)$. *Let τ_R denote the Alexandroff topology on* $\mathrm{uf}(X)$ *generated by R. We call* $\mathrm{ae}(\mathcal{X}) = \langle \mathrm{uf}(X), \tau_R \rangle$ *the* Alexandroff extension *of $\mathcal{X}$.*

We note that if the original topology on X is Alexandroff, then the Alexandroff extension of $\mathcal{X}$ can be obtained by first taking the ultrafilter extension of $\mathcal{X}$ (van Benthem, 1983a) and then taking the corresponding Alexandroff space. Alexandroff extensions of topological spaces turn out to be crucial for obtaining a topological version of the Goldblatt-Thomason theorem.

DEFINITION 5.53 *We say that a class K of topological spaces* reflects Alexandroff extensions *if for each topological space $\mathcal{X}$ we have* $\mathrm{ae}(\mathcal{X}) \in K$ *implies $\mathcal{X} \in K$.*

THEOREM 5.54 (GABELAIA, 2001) *Let K be a class of topological spaces closed under formation of Alexandroff extensions. Then K is modally definable iff it is closed under taking open subspaces, interior images, topological sums, and it reflects Alexandroff extensions.*

Several refinements of Theorem 5.54 and extensions to richer languages can be found in Gabelaia and Sustretov, 2005 and ten Cate et al., 2006. Below we present a number of topo-definable classes of spaces, as well as normal extensions of **S4** being complete with respect to topologically interesting classes of spaces.

The logic of discrete spaces: It is rather easy to see that $\mathcal{X} \models p \to \Box p$ iff every subset of X is open iff $\mathcal{X}$ is discrete. Therefore, $p \to \Box p$ (or equivalently $\Diamond p \to p$) topo-defines the class of discrete spaces.

S5 and clopens: We observe that

$$\begin{aligned}\mathcal{X} \models p \to \Box\Diamond p \quad & \text{iff } A \subseteq \mathrm{Int}(\mathrm{Cl}(A)) \text{ for each } A \subseteq X \\ & \text{iff } \mathrm{Cl}(A) \subseteq \mathrm{Int}(\mathrm{Cl}(A)) \text{ for each } A \subseteq X \\ & \text{iff every closed subset of } X \text{ is open.}\end{aligned}$$

Therefore, $p \to \Box\Diamond p$ (or equivalently $\Diamond p \to \Box\Diamond p$) topo-defines the class of topological spaces in which every closed subset is open.

S4.2 and extremally disconnected spaces: We recall that

$$\mathbf{S4.2} = \mathbf{S4} + (\Diamond\Box p \to \Box\Diamond p).$$

Now observe that

$$\begin{aligned}\mathcal{X} \models \Diamond\Box p \to \Box\Diamond p \quad & \text{iff } \mathrm{Cl}(\mathrm{Int}(A)) \subseteq \mathrm{Int}(\mathrm{Cl}(A)) \text{ for each } A \subseteq X \\ & \text{iff } \mathrm{Cl}(\mathrm{Int}(A)) = \mathrm{Int}(\mathrm{Cl}(\mathrm{Int}(A))) \text{ for each } A \subseteq X \\ & \text{iff the closure of every open subset of } X \text{ is open} \\ & \text{iff } \mathcal{X} \text{ is extremally disconnected.}\end{aligned}$$

Therefore, $\Diamond\Box p \to \Box\Diamond p$ topo-defines the class of extremally disconnected spaces.

S4.1 and filters of dense sets: We recall that

$$\mathbf{S4.1} = \mathbf{S4} + (\Box\Diamond p \to \Diamond\Box p).$$

For a topological space $\mathcal{X}$ let $\mathcal{D}(X)$ be the set of all dense subsets of X. Now $\mathcal{X} \models \Box\Diamond p \to \Diamond\Box p$ iff $\mathrm{Int}(\mathrm{Cl}(A)) \subseteq \mathrm{Cl}(\mathrm{Int}(A))$ for each $A \subseteq X$. As shown in Bezhanishvili et al., 2003, p. 293, Proposition 2.1, the last condition is equivalent to $\mathcal{D}(X)$ being a filter. So, $\Box\Diamond p \to \Diamond\Box p$ topo-defines the class of topological spaces in which $\mathcal{D}(X)$ is a filter.

S4.Grz and hereditarily irresolvable spaces: We recall that

$$\mathbf{S4.Grz} = \mathbf{S4} + \Box(\Box(p \to \Box p) \to p) \to \Box p.$$

We also recall that a space $\mathcal{X}$ is *resolvable* if it can be represented as the union of two disjoint dense subsets, that it is *irresolvable* if it is not resolvable, that it is *hereditarily irresolvable* if every subspace of $\mathcal{X}$ is irresolvable, and that it is *scattered* if every subspace of $\mathcal{X}$ has an isolated point.

For each $A \subseteq X$ let $\rho(A) = A \cap \mathrm{Cl}(\mathrm{Cl}(A) - A)$. Esakia, 1981 observed that $\mathcal{X} \models \Box(\Box(p \to \Box p) \to p) \to \Box p$ iff $A \subseteq \mathrm{Cl}(A - \rho(A))$ for each $A \subseteq X$. As shown in Bezhanishvili et al., 2003, p. 295, Theorem 2.4, the last condition is equivalent to $\mathcal{X}$ being hereditarily irresolvable. Therefore, $\Box(\Box(p \to \Box p) \to p) \to \Box p$ topo-defines the class of hereditarily irresolvable spaces. Now since **S4.Grz** is complete with respect to its relational semantics and since for an Alexandroff space $\mathcal{X}_{\mathfrak{F}} = \langle X, \tau_R \rangle$ the notions of hereditarily irresolvable and scattered coincide with each other and with the notion of $\mathfrak{F}$

Figure 5.14. The 2-fork frame.

having no infinite ascending chains (Gabelaia, 1999), we obtain that **S4.Grz** is the logic of hereditarily irresolvable spaces, and also the logic of scattered spaces. As scattered spaces are a proper subclass of hereditarily irresolvable spaces (Bezhanishvili et al., 2003), they are *not* topo-definable. Moreover, since **S4.Grz** is also the logic of ordinals (Abashidze and Esakia, 1987), ordinals are *not* topo-definable either.

Euclidean hierarchy: Corollary 5.48 shows that the logic of boolean combinations of countable unions of convex subsets of $\mathbb{R}$ is already **S4**. The logic becomes much stronger, however, if we restrict our attention to finite unions of convex subsets of $\mathbb{R}$.

We call a subset of $\mathbb{R}$ *serial* if it is a finite union of convex subsets of $\mathbb{R}$. Let $S(\mathbb{R})$ denote the family of serial subsets of $\mathbb{R}$. Unlike the countable unions of convex subsets of $\mathbb{R}$, $S(\mathbb{R})$ does form a boolean algebra. We call a valuation ν on $\mathbb{R}$ *serial* if $\nu(p) \in S(\mathbb{R})$ for each propositional letter p. We call a formula φ *s-true* if it is true in $\mathbb{R}$ under a serial valuation, and we call φ *s-valid* if φ is s-true for each serial valuation. Let $L(S) = \{\varphi : \varphi \text{ is s-valid}\}$. It is easy to see that $L(S)$ is a normal extension of **S4**, we refer to as *the logic of serial subsets of* $\mathbb{R}$. The following theorem was first established in Aiello et al., 2003 (see also van Benthem et al., 2003):

THEOREM 5.55 *$L(S)$ is the logic of the 2-fork frame $\mathfrak{F}$ shown in Fig. 5.14.*

For $n \geq 2$, we call $X \subseteq \mathbb{R}^n$ *hyper-rectangular convex* if $X = X_1 \times \cdots \times X_n$, where all the X_i's are convex subsets of $\mathbb{R}$. We also call $X \subseteq \mathbb{R}^n$ *n-chequered* if it is a finite union of hyper-rectangular convex subsets of $\mathbb{R}^n$. Let $CH(\mathbb{R}^n)$ denote the set of all n-chequered subsets of $\mathbb{R}^n$. As with $S(\mathbb{R})$, we have that $CH(\mathbb{R}^n)$ forms a boolean algebra. We call a valuation ν on $\mathbb{R}^n$ *n-chequered* if $\nu(p) \in CH(\mathbb{R}^n)$ for each propositional letter p. We call a formula φ *n-true* if it is true in $\mathbb{R}^n$ under an n-chequered valuation, and we call φ *n-valid* if φ is n-true for each n-chequered valuation. Let $L_n = \{\varphi : \varphi \text{ is n-valid}\}$. As with $L(S)$, we have that L_n is a normal extension of **S4**, we refer to as *the logic of n-chequered subsets of* $\mathbb{R}^n$. Moreover, the logics form a decreasing chain:

$$L(S) = L_1 \supset L_2 \supset L_3 \supset \cdots \supset L_n \supset \ldots \ .$$

Let $\mathfrak{F}^n$ denote the Cartesian product of the 2-fork frame $\mathfrak{F}$ with itself n-times. The following theorem was proved in van Benthem et al., 2003:

THEOREM 5.56 *For $n \geq 2$ we have that L_n is the logic of $\mathfrak{F}^n$.*

In particular, the logic of chequered subsets of the real plane coincides with the logic of $\mathfrak{F}^2$. An illustration of $\mathfrak{F}^2$ is given in Fig. 5.15.

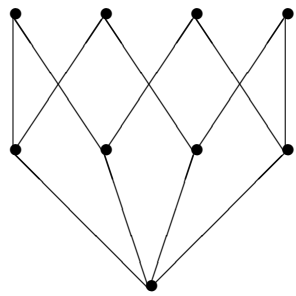


Figure 5.15. $\mathfrak{F}^2$.

We call $X \subseteq \mathbb{R}^\infty$ *∞-rectangular convex* if $X = \prod_{i=1}^{\infty} X_i$, where each X_i is a convex subset of $\mathbb{R}$, and all but finitely many of X_i's are equal to either $\mathbb{R}$ or $\emptyset$. We call $X \subseteq \mathbb{R}^\infty$ *∞-chequered* if it is a finite union of ∞-rectangular convex subsets of $\mathbb{R}^\infty$. Let $CH(\mathbb{R}^\infty)$ denote the set of ∞-chequered subsets of $\mathbb{R}^\infty$. As with each $CH(\mathbb{R}^n)$, we have that $CH(\mathbb{R}^\infty)$ forms a boolean algebra. We call a valuation ν on $\mathbb{R}^\infty$ *∞-chequered* if $\nu(p) \in CH(\mathbb{R}^\infty)$ for each propositional letter p. We call a formula φ *∞-true* if it is true in $\mathbb{R}^\infty$ under a ∞-chequered valuation, and we call φ *∞-valid* if φ is ∞-true for each ∞-chequered valuation. Let $L_\infty = \{\varphi : \varphi \text{ is } \infty\text{-valid}\}$. It is easy to see that L_∞ is a normal extension of **S4**, we refer to as *the logic of ∞-chequered subsets of $\mathbb{R}^\infty$*. The following theorem can be found in van Benthem et al., 2003:

THEOREM 5.57 $L_\infty = \bigcap L_n$.

To summarize, we obtained the logic $L(S)$ of serial subsets of the real line $\mathbb{R}$, as well as its natural generalizations—the logics L_n of sufficiently well-behaved n-chequered subsets of n-dimensional Euclidean spaces $\mathbb{R}^n$. Unlike the full modal logic of each Euclidean space $\mathbb{R}^n$, which coincides with **S4**, all logics L_n are different, forming a decreasing chain converging to the logic L_∞ of ∞-chequered subsets of $\mathbb{R}^\infty$. This provides us with a sort of *Euclidean hierarchy* in modal logic. It has been suggested by Litak, 2004 that L_∞ may be closely related to the quite differently motivated 'Logic of Finite Problems'

first defined by Medvedev. Recently it was shown in Fontaine, 2006 that the Medvedev logic is not finitely axiomatizable over L_∞.

REMARK 5.58 As was pointed out to us by David Gabelaia, $L(S)$ coincides with the logic of the *digital line* (also known as the *Khalimsky line*), while L_2 coincides with the logic of the *digital plane* (also known as the *Khalimsky plane*). Whether this correspondence extends to higher dimensions remains an open problem. For the definition and basic results on digital topologies we refer to Ch. 12.

3. Modal logic and topology: further directions

In Sec. 2 we were chiefly concerned with the interpretation of $\Diamond$ as closure, the resulting logic **S4**, and the landscape of spatial logics over **S4**. But this classical picture is not all there is to modal logic and topology. We already noticed that many important topological properties are not expressible in the basic modal language $\mathcal{L}$. In this section we discuss several different ways of increasing the expressive power of $\mathcal{L}$, viz. modalizing the derivative operator, product constructions on topological spaces, extended modal languages, and epistemic interpretations of topological models. There is no single story-line here, but together these topics show the liveliness of current research.

3.1 $\Diamond$ as derivative

There are at least two natural ways to increase the expressive power of $\mathcal{L}$. One is to add new modal operators to $\mathcal{L}$, and the other is to interpret the modal $\Diamond$ as a topological operator that is more expressive than the closure operator. We consider adding new modal operators to $\mathcal{L}$ in Sec. 3.2 and 3.3. Here we outline some of the consequences of interpreting $\Diamond$ as derivative (first suggested by McKinsey and Tarski, 1944). Since $\mathrm{Cl}(A) = A \cup d(A)$ for each $A \subseteq X$, the derivative operator is more expressive than the closure operator. We recall that $x \in d(A)$ iff $A \cap (U - \{x\}) \neq \emptyset$ for each open neighborhood U of x, that the co-derivative of A is $t(A) = X - d(X - A)$, and that $x \in t(A)$ iff there exists an open neighborhood U of x such that $U \subseteq A \cup \{x\}$.

Let $M = \langle \mathcal{X}, \nu \rangle$ be a topo-model. We define when a formula φ is *d-true at a point* $x \in X$ by induction on the length of φ:

- $x \models_d p$ iff $x \in \nu(p)$;
- $x \models_d \neg\varphi$ iff not $x \models_d \varphi$;
- $x \models_d \varphi \wedge \psi$ iff $x \models_d \varphi$ and $x \models_d \psi$;
- $x \models_d \Box\varphi$ iff $\exists U \in \tau(x \in U \;\&\; \forall y \in U - \{x\} \;\; y \models_d \varphi)$;

and hence also

- $x \models_d \Diamond\varphi$ iff $\forall U \in \tau(x \in U \to \exists y \in U - \{x\} : y \models_d \varphi)$.

We say that φ is *d-true* in $M = \langle \mathcal{X}, \nu \rangle$ if φ is d-true at every $x \in X$. We say that φ is *d-valid* in $\mathcal{X}$ if φ is d-true in every model based on $\mathcal{X}$. Finally, we say that φ is *d-valid* in a class of topological spaces if φ is d-valid in every member of the class.

EXAMPLE 5.59

1 We show that $(p \wedge \Box p) \to \Box\Box p$ is d-valid in **Top**. Let $\mathcal{X} \in$ **Top**, $M = \langle \mathcal{X}, \nu \rangle$ be a topo-model, and $x \models_d p \wedge \Box p$ for $x \in X$. Then $x \models_d p$ and there exists an open neighborhood U of x such that $y \models_d p$ for each $y \in U - \{x\}$. Therefore, $y \models_d p$ for each $y \in U$. But then $y \models_d \Box p$ for each $y \in U$, implying that $x \models_d \Box\Box p$.

2 That $\Box(p \to q) \to (\Box p \to \Box q)$ is d-valid in **Top** and that the necessitation rule preserves d-validity can be proved similarly to Example 5.19.

DEFINITION 5.60 *Let* **wK4** *denote the modal logic* $\mathbf{K} + (p \wedge \Box p) \to \Box\Box p$. *Obviously* **wK4** *is weaker than* **K4**, *and we call* **wK4** weak **K4**.

It follows that the modal logic **wK4** is sound with respect to d-semantics. In fact, as shown in Esakia, 2001, **wK4** is also complete with respect to d-semantics. Below we discuss the connection between relational semantics of **wK4** and d-semantics; then we show that **wK4** is the d-logic of all topological spaces; after that we determine the connection between **S4** and **wK4**; finally, we discuss stronger spatial logics over **wK4**. Most results in this section are taken from Esakia, 2001, Esakia, 2004, Bezhanishvili et al., 2005, Shehtman, 1990 and Shehtman, 2006.

3.1.1 Weak K4.

DEFINITION 5.61 *Let* $\mathfrak{F}$ *be a frame. We call* $\mathfrak{F}$ weakly transitive *if* $\forall w, v, u \in W$ $(wRv \;\&\; vRu \;\&\; w \neq u \to wRu)$.

It is known that **wK4** is sound and complete with respect to the class of all weakly transitive frames. In fact, **wK4** has the finite model property (Esakia, 2001). Because of this, we will sometimes refer to weakly transitive frames as **wK4**-frames. We call a weakly transitive frame $\mathfrak{F} = \langle W, R \rangle$ *rooted* if there exists $r \in W$—called a *root* of $\mathfrak{F}$—such that rRw for every $w \neq r$.

THEOREM 5.62 (ESAKIA, 2001) **wK4** *is complete with respect to finite rooted irreflexive* **wK4**-*frames.*

Proof (Sketch) If $\mathbf{wK4} \nvdash \varphi$, then there exists a finite **wK4**-model $M = \langle \mathfrak{F}, \nu \rangle$ refuting φ. Now we construct a finite irreflexive **wK4**-frame $\mathfrak{G}$ such that $\mathfrak{F}$ is a

p-morphic image of $\mathfrak{G}$. For $\mathfrak{F} = \langle X, R\rangle$ let Y be obtained from X by replacing every reflexive point x of X by a two point set $\{x_1, x_2\}$ disjoint from X. Define S on Y by x_1Sx_2 and x_2Sx_1, ySx_1 and ySx_2 for each yRx, and x_1Sz and x_2Sz for each xRz. It is easy to see that $\mathfrak{G}$ is a finite irreflexive weakly transitive frame. Define $f : Y \to X$ by $f(x_1) = f(x_2) = x$ if $x \in X$ is reflexive, and $f(x) = x$ if $x \in X$ is irreflexive. Then it is routine to check that f is an onto p-morphism. Now define ξ on $\mathfrak{G}$ by $\xi(p) = f^{-1}(\nu(p))$. Then the model $N = \langle \mathfrak{G}, \xi\rangle$ is bisimilar to M. Therefore, N also refutes φ. QED

DEFINITION 5.63 *Let $\mathfrak{F} = \langle X, R\rangle$ be a* **wK4***-frame. We denote by $\overline{\mathfrak{F}} = \langle X, \overline{R}\rangle$ the* reflexive closure *of $\mathfrak{F}$ (that is, $\overline{R}$ is obtained from R by adding* all *reflexive loops), and by $\underline{\mathfrak{F}} = \langle X, \underline{R}\rangle$ the* irreflexive fragment *of $\mathfrak{F}$ (that is, $\underline{R}$ is obtained from R by deleting* all *reflexive loops).*

For a **wK4**-frame $\mathfrak{F}$, it is obvious that $\overline{\mathfrak{F}}$ is an **S4**-frame, and that $\underline{\mathfrak{F}}$ is an irreflexive **wK4**-frame. Moreover, every **wK4**-frame is obtained either from an **S4**-frame by deleting some reflexive loops or from an irreflexive **wK4**-frame by adding some reflexive loops.

Given a **wK4**-frame $\mathfrak{F}$, we view $\overline{\mathfrak{F}}$ as an Alexandroff space. For $A \subseteq X$ we denote the derivative of A in $\overline{\mathfrak{F}}$ by $d_R(A)$. The next series of results is taken from Esakia, 2001.

LEMMA 5.64 *Let $\mathfrak{F}$ be a* **wK4***-frame and $A \subseteq X$. In $\overline{\mathfrak{F}}$ we have $d_R(A) = \underline{R}^{-1}(A)$.*

Proof We observe that

$$\begin{aligned} x \in d_R(A) \quad & \text{iff for each open neighborhood } U \text{ of } x \text{ we have } U \cap (A - \{x\}) \neq \emptyset \\ & \text{iff } \overline{R}(x) \cap (A - \{x\}) \neq \emptyset \\ & \text{iff } \underline{R}(x) \cap A \neq \emptyset \\ & \text{iff } x \in \underline{R}^{-1}(A). \end{aligned}$$

The result follows. QED

Now suppose $\mathcal{X}$ is a topological space. We define R_d on X by xR_dy iff $x \in d(y)$.

LEMMA 5.65 *$\langle X, R_d\rangle$ is an irreflexive* **wK4***-frame.*

Proof That R_d is irreflexive follows from $x \notin d(x)$. To see that R_d is weakly transitive suppose xR_dy, yR_dz, and $x \neq z$. Then $x \in d(y)$, $y \in d(z)$, and $x \neq z$. From $x \in d(y)$ it follows that for each open neighborhood U of x we have $y \in U - \{x\}$; from $y \in d(z)$ it follows that for each open neighborhood V of y we have $z \in V - \{y\}$; and from $x \neq z$ it follows that for each open neighborhood U of x we have $z \in U - \{x, y\} \subseteq U - \{x\}$. Thus, $x \in d(z)$, and so xR_dz. QED

LEMMA 5.66

1 *If $\mathfrak{F}$ is a* **wK4***-frame, then $R_{d_R} \subseteq R$.*

2 *If $\mathfrak{F}$ is an irreflexive* **wK4***-frame, then $R_{d_R} = R$.*

3 *If $\mathcal{X}$ is a topological space, then $R_d^{-1}(A) \subseteq d(A)$.*

4 *If $\mathcal{X}$ is an Alexandroff space, then $R_d^{-1}(A) = d(A)$.*

Proof (1) In a **wK4**-frame $\mathfrak{F}$, $xR_{d_R}y \to x \in d_R(y) \to x \in R^{-1}(y) \to xRy$.
(2) Suppose $\mathfrak{F}$ is an irreflexive **wK4**-frame. Then $xR_{d_R}y \leftrightarrow x \in d_R(y) \leftrightarrow x \in \underline{R}^{-1}(y) \leftrightarrow x \in R^{-1}(y) \leftrightarrow xRy$.
(3) Suppose $\mathcal{X}$ is a topological space. Then $x \in R_d^{-1}(A) \to (\exists y)(xR_dy \& y \in A) \to (\exists y)(x \in d(y) \ \& \ d(y) \subseteq d(A)) \to x \in d(A)$.
(4) Suppose $\mathcal{X}$ is an Alexandroff space. Then $x \in d(A) \leftrightarrow \overline{R_d}(x) \cap (A - \{x\}) \neq \emptyset \leftrightarrow R_d(x) \cap A \neq \emptyset \leftrightarrow x \in R_d^{-1}(A)$. QED

COROLLARY 5.67 *For a nonempty set X, there is a 1-1 correspondence between:*

(i) Alexandroff topologies on X;

(ii) Reflexive and transitive relations on X;

(iii) Irreflexive and weakly transitive relations on X.

It follows that there is a 1-1 correspondence between Alexandroff spaces, **S4**-frames, and irreflexive **wK4**-frames. Now we are in a position to show that **wK4** is the d-logic of all topological spaces.

THEOREM 5.68

1 **wK4** *is the d-logic of all topological spaces.*

2 **wK4** *is the d-logic of all* finite *topological spaces.*

3 **wK4** *has the* effective finite model *property with respect to the class of topological spaces.*

Proof Obviously both (1) and (3) follow from (2). To see (2), Theorem 5.62 implies that **wK4** is complete with respect to finite irreflexive **wK4**-frames. By Corollary 5.67, irreflexive **wK4**-frames correspond to topological spaces. The result follows. QED

3.1.2 Connections between S4 and wK4. There is a close connection between **S4** and **wK4**. For the set $\mathbb{F}$ of formulas of $\mathcal{L}$, we define a *translation* $tr : \mathbb{F} \to \mathbb{F}$ by induction (Boolos, 1993):

- $tr(p) = p$;
- $tr(\varphi \wedge \psi) = tr(\varphi) \wedge tr(\psi)$;
- $tr(\neg\varphi) = \neg tr(\varphi)$;
- $tr(\Box\varphi) = tr(\varphi) \wedge \Box tr(\varphi)$.

DEFINITION 5.69

1 *Let L be a normal extension of* **wK4** *and S be a normal extension of* **S4**. *We say that L and S are* companions *if $S \vdash \varphi$ iff $L \vdash tr(\varphi)$.*

2 *For a normal extension L of* **wK4**, *let $T(L) = \{\varphi \ : \ L \vdash tr(\varphi)\}$.*

LEMMA 5.70

1 *$T(L)$ is a normal extension of* **S4**.

2 *$T(L)$ is a unique companion of L.*

Proof (1) It is easy to verify that

$$\mathbf{wK4} \vdash tr(\Box p \to p), tr(\Box p \to \Box\Box p), tr(\Box(p \to q) \to (\Box p \to \Box q)),$$

and that $T(L)$ is closed under MP and N. So, $T(L)$ is a normal extension of **S4**.

(2) Suppose S is a companion of L. Then $S \vdash \varphi \leftrightarrow L \vdash tr(\varphi) \leftrightarrow T(L) \vdash \varphi$. Therefore, $S = T(L)$. QED

On the other hand, a given normal extension S of **S4** may have many different companions. For example, both **wK4** and **K4** (and all the normal logics in between) are companions of **S4**.

Speaking in terms of relational semantics, if a normal extension L of **wK4** is characterized by a class K of frames, then $T(L)$ is characterized by the class $\overline{K} = \{\overline{\mathfrak{F}} \ : \ \mathfrak{F} \in K\}$, where $\overline{\mathfrak{F}}$ denotes the reflexive closure of $\mathfrak{F}$.

Now we turn to the topological significance of tr. For a class K of topological spaces, let $L_d(K)$ denote the set of formulas of $\mathcal{L}$ that are d-valid in K. Since **wK4** is sound with respect to d-semantics, it is obvious that $L_d(K)$ is a normal extension of **wK4**. We call $L_d(K)$ the *d-logic* of the class K. The next two facts are taken from Bezhanishvili et al., 2005, Lemma 2.1 and Theorem 2.2.

LEMMA 5.71 *Let K be a class of topological spaces and $\mathcal{X} \in K$.*

1 *$\mathcal{X} \models \varphi$ iff $\mathcal{X} \models_d tr(\varphi)$.*

2 $K \models \varphi$ *iff* $K \models_d tr(\varphi)$.

THEOREM 5.72 *If L is a d-logic, then $T(L)$ is topologically complete.*

We point out that $T(L)$ may be topologically complete without L being a d-logic. Thus, the converse of Theorem 5.72 is not true in general. Next we indicate several examples of topologically complete normal extensions of **S4** and **wK4** that are each others companions.

3.1.3 The landscape of spatial logics over wK4. The landscape of spatial logics over **wK4** has been investigated less vigorously than that of spatial logics over **S4**. Nevertheless, there are several interesting results in this direction that we list below.

As we have already seen, **wK4** is the logic of all topological spaces when interpreting $\Diamond$ as derivative. In addition, we will see that **K4** is the logic of all T_d-spaces, that **KD4** is the logic of the Cantor space $\mathbb{C}$, the rational line $\mathbb{Q}$, or more generally, any 0-dimensional dense-in-itself metric separable space, that the d-logic of $\mathbb{R}$ is $\mathbf{KD4G}_2$, and that the d-logic of each $\mathbb{R}^n$, for $n \geq 2$, is $\mathbf{KD4G}_1$.

We say that a class K of topological spaces is *d-definable* if there exists a set of modal formulas Γ such that for each topological space $\mathcal{X}$ we have $\mathcal{X} \in K$ iff $\mathcal{X} \models_d \Gamma$. Since the derivative operator of a topological space is more expressive than the closure operator, topo-definability results will automatically transfer into d-definability results. However, there are d-definable topological properties that are not topo-definable. For example, the class of all T_d-spaces is *not* topo-definable. On the other hand, $\Box p \to \Box\Box p$ d-defines it. Also, the class of dense-in-itself spaces is *not* topo-definable, but it is d-definable by $\Diamond\top$ (or equivalently by $\Box p \to \Diamond p$). Below we present several results in this direction.

K4 and T_d-spaces:

PROPOSITION 5.73 (ESAKIA, 2001) **K4** *is the d-logic of all T_d-spaces.*

Proof Since $\mathcal{X}$ is a T_d-space iff $dd(A) \subseteq d(A)$ for each $A \subseteq X$, we obtain that **K4** is sound with respect to the class of T_d-spaces. To see completeness, recall that **K4** is complete with respect to the class of all (not necessarily finite) irreflexive **K4**-frames (see, e.g., Chagrov and Zakharyaschev, 1997, p. 102, Exercise 3.11). Since each one of these corresponds to a T_d-space, the result follows. QED

Another way to obtain d-completeness of **K4** is to construct a canonical topological model for **K4** similar to the one constructed in Sec. 2.4.2 (Steinsvold, 2005).

The d-logics of $\mathbb{C}$ and $\mathbb{Q}$: Let

$$\mathbf{KD4} = \mathbf{K4} + \Diamond\top.$$

For a topological space $\mathcal{X}$ we have $\mathcal{X} \models_d \Diamond\top$ iff $d(X) = X$ iff $\mathcal{X}$ is dense-in-itself. Consequently, since both $\mathbb{C}$ and $\mathbb{Q}$ are dense-in-itself T_d-spaces, we have that $\mathbb{C}, \mathbb{Q} \models_d \mathbf{KD4}$. Moreover, as shown in Shehtman, 1990, Theorem 29, **KD4** is the d-logic of any 0-dimensional dense-in-itself metric separable space. As an immediate consequence we obtain that $\mathbf{KD4} = L_d(\mathbb{C}) = L_d(\mathbb{Q})$.

The d-logics of Euclidean spaces: The situation here is different from that of **S4**. Indeed, we have that the d-logic of $\mathbb{C}$ and $\mathbb{Q}$ is different from the d-logic of $\mathbb{R}$, and that the d-logic of $\mathbb{R}$ is different from the d-logic of $\mathbb{R}^n$ for each $n \geq 2$. Let G_1 denote the formula

$$(\Diamond p \wedge \Diamond\neg p) \to \Diamond((p \wedge \Diamond\neg p) \vee (\neg p \wedge \Diamond p)),$$

and $\mathbf{KD4G}_1$ denote the logic obtained from **KD4** by postulating the formula G_1. Also let $Q_1 = p_1 \wedge \neg p_2 \wedge \neg p_3$, $Q_2 = \neg p_1 \wedge p_2 \wedge \neg p_3$, $Q_3 = \neg p_1 \wedge \neg p_2 \wedge p_3$, and G_2 denote the formula

$$\Box((Q_1 \wedge \Box Q_1) \vee (Q_2 \wedge \Box Q_2) \vee (Q_3 \wedge \Box Q_3)) \to (\Box\neg Q_1 \vee \Box\neg Q_2 \vee \Box\neg Q_3).$$

We let $\mathbf{KD4G}_2$ denote the logic obtained from **KD4** by postulating the formula G_2. Then it follows from Shehtman, 1990, Shehtman, 2006 that $\mathbf{KD4G}_1 = L_d(\mathbb{R}^n)$ for each $n \geq 2$, and that $\mathbf{KD4G}_2 = L_d(\mathbb{R})$.

GL and scattered spaces: Recall that the Gödel-Löb provability logic **GL** is obtained by adding to **K** the following Löb formula $\Box(\Box p \to p) \to \Box p$. As we already saw in Sec. 2.6, the class of scattered spaces is not topo-definable. On the other hand, as shown in Esakia, 1981, $\mathcal{X} \models_d \Box(\Box p \to p) \to \Box p$ iff $\mathcal{X}$ is scattered. Therefore, $\Box(\Box p \to p) \to \Box p$ d-defines the class of scattered spaces. Now as **GL** is the d-logic of ordinal spaces (Abashidze, 1987; Blass, 1990), we obtain that **GL** is the d-logic of both scattered spaces and ordinal spaces. Since the class of ordinal spaces is a proper subclass of the class of scattered spaces, it follows that the class of ordinal spaces is neither d-definable nor topo-definable.

K4.Grz and hereditarily irresolvable spaces: As we already saw in Sec. 2.6, $\Box(\Box(p \to \Box p) \to p) \to \Box p$ topo-defines the class of hereditarily irresolvable spaces. Consequently, the class of hereditarily irresolvable spaces is also d-definable. Interestingly enough, the same axiom d-defines the class of hereditarily irresolvable spaces (Esakia, 2002). Moreover, **K4.Grz** is the d-logic of hereditarily irresolvable spaces (Gabelaia, 2004). Both **GL** and **K4.Grz** are companions of **S4.Grz**, but as shown in Esakia, 2002, **K4.Grz** is the least companion of **S4.Grz**.

Further results on d-definability and d-completeness can be found in Bezhanishvili et al., 2005.

3.2 Product logics

Throughout the 1990's, combinations of logics and constructions merging their models have come up in practice, and also as a new theme for mathematical theory. In particular, *products* of relational models have been studied extensively in Gabbay and Shehtman, 1998 for their uses in combining information along different dimensions, and the behavior of the matching modal logics is well-known (Gabbay et al., 2003). This section is about products of topological spaces as a generalization of this methodology. In particular, we describe two 'horizontal' and 'vertical' topologies along with the standard product topology. For each of the three topologies on the product we introduce a modal box in our language and give axiomatizations of the resulting logics. The material presented here is taken from van Benthem et al., 2005.

3.2.1 Products of topologies. Let $\mathcal{X} = \langle X, \eta \rangle$ and $\mathcal{Y} = \langle Y, \theta \rangle$ be two topological spaces. Recall that the *standard product topology* τ on $X \times Y$ is defined by letting the sets $U \times V$ form a basis for τ, where U is open in $\mathcal{X}$ and V is open in $\mathcal{Y}$. We define two additional one-dimensional topologies on $X \times Y$ by 'lifting' the topologies of the components.

DEFINITION 5.74 *Suppose $A \subseteq X \times Y$. We say that A is* horizontally open (H-open) *if for any $(x, y) \in A$ there exists $U \in \eta$ such that $x \in U$ and $U \times \{y\} \subseteq A$. Similarly, we say that A is* vertically open (V-open) *if for any $(x, y) \in A$ there exists $V \in \theta$ such that $y \in V$ and $\{x\} \times V \subseteq A$. If A is both H- and V-open, then we call it* HV-open.

The H-closed, V-closed and HV-closed sets are defined similarly. Let τ_1 denote the set of all H-open subsets of $X \times Y$ and τ_2 denote the set of all V-open subsets of $X \times Y$. It is easy to verify that both τ_1 and τ_2 form topologies on $X \times Y$.

DEFINITION 5.75 *We call τ_1 the* horizontal topology *and τ_2 the* vertical topology.

REMARK 5.76 A set open in the standard product topology is both horizontally and vertically open. That is $\tau \subseteq \tau_1$ and $\tau \subseteq \tau_2$. However, the converse inclusions do not hold in general.

The interpretation of the modal operators $\Box_1$ and $\Box_2$ of $\mathcal{L}_{\Box_1 \Box_2}$ in $\langle X \times Y, \tau_1, \tau_2 \rangle$ is as expected:

$$(x, y) \models \Box_1 \varphi \text{ iff } (\exists U \in \tau_1)((x, y) \in U \text{ and } \forall (x', y') \in U \ (x', y') \models \varphi)$$

$$(x, y) \models \Box_2 \varphi \text{ iff } (\exists V \in \tau_2)((x, y) \in V \text{ and } \forall (x', y') \in V \ (x', y') \models \varphi).$$

The modalities $\Diamond_1$ and $\Diamond_2$ are defined dually. Furthermore, all the usual notions such as satisfiability and validity generalize naturally to this new language.

There are some similarities and differences between products of frames and of topological spaces. To see the similarities, let $\mathfrak{F} = \langle W, R\rangle$ and $\mathfrak{F}' = \langle W', R'\rangle$ be **S4**-frames, and let $\mathfrak{F} \times \mathfrak{F}' = \langle W \times W', R_1, R_2\rangle$ be their product. Then τ_{R_1} and τ_{R_2} are precisely the horizontal and vertical topologies on the product space $W \times W'$. This shows that our topological product construction is a faithful generalization of the usual product construction for frames. Thus, whenever topological spaces $\mathcal{X}$ and $\mathcal{Y}$ are representable as **S4**-frames (are Alexandroff), then the horizontal and vertical topologies on their product $X \times Y$ can be defined from the horizontal and vertical relations on the product of these frames. In other words, our topological setting generalizes the case for products of frames. To see the differences, we point out that both com and chr, while valid on products of frames, can be refuted on topological products. Below we exhibit their failure on $\mathbb{R} \times \mathbb{R}$.

(a) Failure of com: Let

$$\nu(p) = (\bigcup_{x\in(-1,0)} \{x\} \times (x, -x)) \cup (\{0\} \times (-1, 1)) \cup (\bigcup_{x\in(0,1)} \{x\} \times (-x, x))$$

(see Fig. 5.16a). Then there is a basic horizontal open $(-1, 1) \times \{0\}$ such that $(0, 0)$ is in it and every point in $(-1, 1) \times \{0\}$ sits in a vertically open subset of $\nu(p)$. Thus, $\Box_1\Box_2 p$ is true at $(0, 0)$. On the other hand, there is no vertical open containing $(0, 0)$ in which every point sits inside a horizontally open subset of $\nu(p)$, implying that $\Box_2\Box_1 p$ is false at $(0, 0)$.

(b) Failure of chr: Let $\nu(p) = \bigcup\{\{\frac{1}{n}\} \times (-\frac{1}{n}, \frac{1}{n}) : n \geq 1\}$ (see Fig. 5.16b). Then in any basic horizontal open around $(0, 0)$ there is a point that sits in a basic vertical open in which p is true everywhere. Thus, $\Diamond_1\Box_2 p$ is true at $(0, 0)$. On the other hand, since the horizontal closure of $\nu(p)$ is $\nu(p) \cup \{(0, 0)\}$ and since the vertical interior of $\nu(p) \cup \{(0, 0)\}$ is $\nu(p)$, we have that $\Box_2\Diamond_1 p$ is false at $(0, 0)$.

These counterexamples on $\mathbb{R} \times \mathbb{R}$ are not accidental. van Benthem et al., 2005, Sec. 4 shows when they can be reproduced in products of arbitrary topological spaces.

3.2.2 Completeness for products. Our main goal here is to show that the logic of all products of topological spaces is $\mathbf{S4} \oplus \mathbf{S4}$. In fact, we show that $\mathbf{S4} \oplus \mathbf{S4}$ is the logic of $\mathbb{Q} \times \mathbb{Q}$.

THEOREM 5.77 $\mathbf{S4} \oplus \mathbf{S4}$ *is the logic of* $\mathbb{Q} \times \mathbb{Q}$.

Proof As follows from Proposition 5.17, $\mathbf{S4} \oplus \mathbf{S4}$ is complete with respect to the infinite quaternary tree $\mathcal{T}_{2,2} = \langle W, R_1, R_2\rangle$. We view $\mathcal{T}_{2,2}$ as equipped with

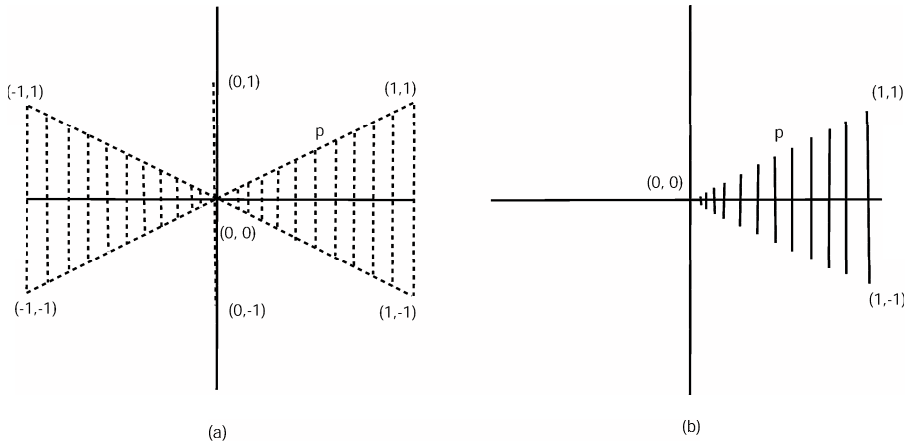


Figure 5.16. Counterexamples of com and chr on $\mathbb{R} \times \mathbb{R}$.

two Alexandroff topologies defined from R_1 and R_2. To prove completeness of $\mathbf{S4} \oplus \mathbf{S4}$ with respect to $\mathcal{CQ} \times \mathcal{CQ}$ we take the X constructed in the proof of Theorem 5.36, define recursively an HV-open subspace Y of $X \times X$ and an interior map g from Y onto $\mathcal{T}_{2,2}$ with respect to both topologies: this allows us to transfer counterexamples from $\mathcal{T}_{2,2}$ to Y, then from Y to $X \times X$, and finally from $X \times X$ to $\mathcal{CQ} \times \mathcal{CQ}$.

Let $Y = \bigcup_{n \in \omega} Y_n$ where $Y_0 = \{(0,0)\}$ and

$$Y_{n+1} \;=\; Y_n \cup \{(x - \frac{1}{3^n}, y), (x + \frac{1}{3^n}, y), (x, y - \frac{1}{3^n}), (x, y + \frac{1}{3^n}) \;:\; (x,y) \in Y_n\}$$

CLAIM 5.78 *Y is an HV-open subspace of $X \times X$.*

Proof Let $(x,y) \in Y$. Then $x \in (x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}}) \subseteq X$. Therefore, $(x,y) \in (x - \frac{1}{3^{n_x}}, x + \frac{1}{3^{n_x}}) \times \{y\} \subseteq Y$. Thus, Y is an H-open subspace of $X \times X$. That Y is a V-open subspace of $X \times X$ is proved symmetrically. QED

A similar argument to that used in Theorem 5.36 shows that for each $(x,y) \in Y$ such that $(x,y) \neq (0,0)$ there exists $n_{(x,y)}$ with $(x,y) \in Y_{n_{(x,y)}}$ and $(x,y) \notin Y_{n_{(x,y)}-1}$, and that there is a unique $(u,v) \in Y_{n_{(x,y)}-1}$ such that $(x,y) = (u \pm \frac{1}{3^{n_{(x,y)}-1}}, v)$ or $(x,y) = (u, v \pm \frac{1}{3^{n_{(x,y)}-1}})$.

We define g from Y onto $\mathcal{T}_{2,2}$ by recursion (cf. Fig. 5.17): If $(x,y) = (0,0)$ then we let $g(0,0)$ be the root r of $\mathcal{T}_{2,2}$; if $(x,y) \neq (0,0)$ then $(x,y) = (u \pm \frac{1}{3^{n_{(x,y)}-1}}, v)$ or $(x,y) = (u, v \pm \frac{1}{3^{n_{(x,y)}-1}})$ for a unique $(u,v) \in Y_{n_{(x,y)}-1}$,

and we let

$$g(x,y) = \begin{cases} \text{the immediate left } R_1\text{-successor of } g(u,v) \\ \qquad \text{if } (x,y) = (u - \frac{1}{3^{n(x,y)-1}}, v) \\ \text{the immediate right } R_1\text{-successor of } g(u,v) \\ \qquad \text{if } (x,y) = (u + \frac{1}{3^{n(x,y)-1}}, v) \\ \text{the immediate left } R_2\text{-successor of } g(u,v) \\ \qquad \text{if } (x,y) = (u, v - \frac{1}{3^{n(x,y)-1}}) \\ \text{the immediate right } R_2\text{-successor of } g(u,v) \\ \qquad \text{if } (x,y) = (u, v + \frac{1}{3^{n(x,y)-1}}) \end{cases}$$

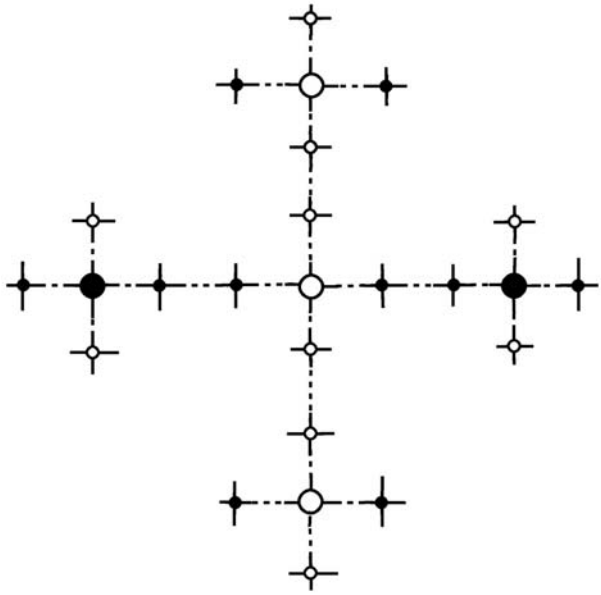


Figure 5.17. The first stages of the labelling in the completeness proof for $\mathbf{S4} \oplus \mathbf{S4}$: $(0,0)$ is labelled by the root r of $\mathcal{T}_{2,2}$, $(-1,0)$ is labelled by the immediate left R_1-successor of r, $(1,0)$ is labelled by the immediate right R_1-successor of r, $(0,-1)$ is labelled by the immediate left R_2-successor of r, $(0,1)$ is labelled by the immediate right R_2-successor of r, and so on.

CLAIM 5.79 *g is an interior map with respect to both topologies.*

Proof Let τ_1 and τ_2 denote the restrictions of the horizontal and vertical topologies of $X \times X$ to Y, respectively. We prove that g is an interior map with respect to τ_1. That g is interior with respect to τ_2 is proved symmetrically. We observe that

$$\{(x - \frac{1}{3^{n(x,y)}}, x + \frac{1}{3^{n(x,y)}}) \times \{y\} \ : \ (x,y) \in Y\}$$

forms a basis for τ_1. We also recall that a basis for the Alexandroff topology on $T_{2,2}$ defined from R_1 is $\mathcal{B}_1 = \{B^1_t\}_{t \in T_{2,2}}$ where $B^1_t = \{s \in T_{2,2} : tR_1 s\}$.

To see that g is open, let $(x - \frac{1}{3^{n(x,y)}}, x + \frac{1}{3^{n(x,y)}}) \times \{y\}$ be a basic open for τ_1. Then the same argument as that used in Claim 5.38 guarantees that $g((x - \frac{1}{3^{n(x,y)}}, x + \frac{1}{3^{n(x,y)}}) \times \{y\}) = B^1_{g(x,y)}$. Thus, g is open. To see that g is continuous, it suffices to show that for each $t \in T_{2,2}$, the g-inverse image of B^1_t belongs to τ_1. Let $(x, y) \in g^{-1}(B^1_t)$. Then $tR_1 g(x, y)$. So $g((x - \frac{1}{3^{n(x,y)}}, x + \frac{1}{3^{n(x,y)}}) \times \{y\}) = B^1_{g(x,y)} \subseteq B^1_t$. Thus, there exists an open neighborhood $U = (x - \frac{1}{3^{n(x,y)}}, x + \frac{1}{3^{n(x,y)}}) \times \{y\}$ of (x, y) such that $U \subseteq g^{-1}(B^1_t)$, implying that g is continuous. QED

To complete the proof, if $\mathbf{S4} \oplus \mathbf{S4} \not\vdash \varphi$, then there is a valuation ν on $\mathcal{T}_{2,2}$ such that $\langle \mathcal{T}_{2,2}, \nu \rangle, r \not\models \varphi$. Define a valuation ξ on Y by $\xi(p) = g^{-1}(\nu(p))$. Since g is an interior map with respect to both topologies and $g(0, 0) = r$, we have that $\langle Y, \xi \rangle, (0, 0) \not\models \varphi$. Now since Y is an HV-open subset of $X \times X$, we obtain that φ is refutable on $X \times X$. Finally, Theorem 5.36 implies that X is homeomorphic to $\mathbb{Q}$. Therefore, $X \times X$ is both horizontally and vertically homeomorphic to $\mathbb{Q} \times \mathbb{Q}$, and hence φ is also refutable on $\mathbb{Q} \times \mathbb{Q}$. QED

COROLLARY 5.80 $\mathbf{S4} \oplus \mathbf{S4}$ *is the logic of products of arbitrary topologies.*

It follows that the logic of products of arbitrary topologies is decidable and has a $PSPACE$-complete satisfiability problem (Spaan, 1993). This stands in contrast to the satisfiability problem for $\mathbf{S4} \times \mathbf{S4}$, which turned out to be undecidable (Gabelaia et al., 2005).

3.2.3 Adding standard product interior. So far we only focused on the horizontal and vertical topologies on the product space, by analogy to products of relational structures. However, unlike products of relational structures, the standard product topology is not definable in terms of the horizontal and vertical topologies (van Benthem et al., 2005, Sec. 3). Therefore, it is only natural to add an extra modal operator $\Box$ to the language $\mathcal{L}_{\Box_1 \Box_2}$ with the intended interpretation as the interior operator of the standard product topology.

For two topological spaces $\mathcal{X} = \langle X, \eta \rangle$ and $\mathcal{Y} = \langle Y, \theta \rangle$, we consider the product $\langle X \times Y, \tau, \tau_1, \tau_2 \rangle$ with three topologies: the standard product topology τ, the horizontal topology τ_1, and the vertical topology τ_2. Then $\Box$ is interpreted as:

$$(x, y) \models \Box\varphi \quad \text{iff} \quad \exists U \in \eta \text{ and } \exists V \in \theta : U \times V \models \varphi.$$

Since $\tau \subseteq \tau_1 \cap \tau_2$, the modal principle

$$\Box p \to \Box_1 p \wedge \Box_2 p$$

is valid in product spaces. Let **TPL** denote the logic in the new language $\mathcal{L}_{\Box,\Box_1,\Box_2}$ that contains all the axioms of $\mathbf{S4} \oplus \mathbf{S4} \oplus \mathbf{S4}$ plus the axiom $\Box p \to \Box_1 p \wedge \Box_2 p$. We call **TPL** the *topological product logic*. The main significance of **TPL** is that in the language $\mathcal{L}_{\Box,\Box_1,\Box_2}$ it is the logic of products of arbitrary topologies. This can be proved by generalizing the completeness of $\mathbf{S4} \oplus \mathbf{S4}$ with respect to $\mathcal{CQ} \times \mathcal{CQ}$ for this new case. As a result, we obtain that $\mathbf{TPL}$ is complete with respect to $\mathcal{CQ} \times \mathcal{CQ}$, hence is the logic of products of arbitrary topologies (with horizontal, vertical, and standard product topologies). For the details of the proof see van Benthem et al., 2005, Sec. 6.

3.3 Extended modal languages

In modern modal logic one conspicuous trend is design of languages whose expressive power matches the intended application rather than merely working with some formalism 'because our forefathers did it'. All this is subject to the balance, matching expressive power with low complexity, preferably decidable (de Rijke, 1993; van Benthem, 1991b). For instance, to make global assertions about a model, one adds a 'universal modality' $U\phi$ saying that ϕ is true in all worlds. The same move makes sense for space. Topological relations not captured by the basic modal language can be safely expressed by adding appropriate new modal operators. We have entered the realm of extended or 'hybrid' modal languages (Areces and ten Cate, 2006).

3.3.1 Universal modalities and global properties. The basic language $\mathcal{L}$ interpreted on topological spaces has a 'local' view of the world. A global perspective comes from the addition of the universal modality that expresses accessibility to any point (Goranko and Passy, 1992). Universal modalities were brought to the spatial reasoning community in Bennett, 1995. For this purpose one adds:

$$\begin{array}{lll} M, x \models E\varphi & \text{iff} & \exists y \in X : M, y \models \varphi \\ M, x \models U\varphi & \text{iff} & \forall y \in X \; M, y \models \varphi. \end{array}$$

More systematically the relevant new valid principles are those of **S5**:

(Dual)	$Ep \leftrightarrow \neg U \neg p$
(K)	$U(p \to q) \to (Up \to Uq)$
(T)	$Up \to p$
(4)	$Up \to UUp$
(B)	$p \to UEp.$

In addition, the following 'connecting' principle is part of the axioms:

$$\Diamond p \to Ep.$$

Using these principles, the enriched language $\mathcal{L}_u$ allows a straightforward normal form:

PROPOSITION 5.81 *Every formula of $\mathcal{L}_u$ is equivalent to one without nested occurrences of E, U.*

The definition of topo-bisimulation extends straightforwardly. It merely demands that topo-bisimulations be *total* relations.

THEOREM 5.82 (AIELLO AND VAN BENTHEM, 2002A)

- *Extended modal formulas in $\mathcal{L}_u$ are invariant under total topo-bisimulations.*
- *Finite $\mathcal{L}_u$-modally equivalent models are totally topo-bisimilar.*

In the topological setting, fragments of this language can also be relevant. E.g., a continuous map has only one of the zig-zag clauses of topo-bisimulation. Now, consider 'existential' modal formulas constructed using only atomic formulas and their negations, $\wedge, \vee, \Box, E$, and U.

COROLLARY 5.83 (AIELLO AND VAN BENTHEM, 2002A) *Let the simulation $\rightharpoonup$ run from M to M' with $x \rightharpoonup x'$. Then, for any existential modal formula φ, $M, x \models \varphi$ only if $M', x' \models \varphi$. In words, existential modal formulas are preserved under simulations.*

The language $\mathcal{L}_u$ is more expressive than the basic modal language $\mathcal{L}$. Indeed, as we already saw in Sec. 2.6, connectedness of a topological space is not expressible in $\mathcal{L}$. However, as shown in Shehtman, 1999, it *is* expressible in $\mathcal{L}_u$ by the formula:

$$U(\Diamond p \to \Box p) \to Up \vee U\neg p. \tag{5.1}$$

In fact, it was shown in Shehtman, 1999 that (5.1) axiomatizes any connected dense-in-itself metric separable space, which is a generalization of the McKinsey-Tarski theorem. Another generalization is in Bezhanishvili and Gehrke, 2005, where it is shown that (5.1) axiomatizes the boolean combinations of countable unions of convex subsets of the real line.

By encoding a fragment of the Region Connection Calculus (RCC5) (Randell et al., 1992) in the language $\mathcal{L}_u$, Bennett showed the power of the language in expressing spatial arrangement of regions. The relevant elementary relations between regions that one can express are those of parthood and connectedness. The encoding is reported in Fig. 5.18, which is the basis for the appropriate calculus in computer science and AI. For details, cf. Ch. 3.

We have just given the first steps here of a much longer ladder. Much stronger hybrid languages, all the way up to first-order languages for topological models

RCC5	$\mathcal{L}_u$	Description
$\mathtt{DC}(A, B)$	$\neg E(A \wedge B)$	A is disconnected from B
$\mathtt{EC}(A, B)$	$E(\Diamond A \wedge \Diamond B) \wedge \neg E(\Box A \wedge \Box B)$	A and B are externally connected
$\mathtt{P}(A, B)$	$U(A \to B)$	A is part of B
$\mathtt{EQ}(A, B)$	$U(A \leftrightarrow B)$	A and B are equal

Figure 5.18. Expressing RCC5 relations via $\mathcal{L}_u$.

from abstract model theory, are found in ten Cate et al., 2006. Still richer descriptions arise in modal predicate logics, which quantify over individuals and refer to arbitrary predicates. Cf. Rasiowa and Sikorski, 1963; Moerdijk 1982 and Awodey and Kirshida, 2006 for such logics which are complete for topological and sheaf semantics.

3.3.2 Temporal operators and boundaries. Another kind of extension of the modal language of topology comes from temporal logic. Consider the 'Until' language of Kamp, 1968. Abstracting from linear temporal behavior gives a natural notion of spatial 'Until', describing truth in a neighborhood up to some 'fence' in topological models:

$M, x \models \varphi \mathcal{U} \psi$ iff there exists an open neighborhood U of x such that $\forall y \in U$ we have $\varphi(y)$ and $\forall z$ on the boundary of U we have $\psi(z)$.

Here the boundary is definable in the earlier modal language:

$$\text{boundary}(U) = \Diamond U \wedge \Diamond \neg U.$$

As in temporal logic, this operator can define various further notions of interest. This richer language still has topo-bisimulations in line with the proposals in Kurtonina and de Rijke, 1997 for dealing with the $\exists\forall$-complexity of 'Until'.

Borrowing from temporal logic is an interesting phenomenon per se. Many temporal principles valid in $\mathbb{R}$ survive the move to the spatial interpretation. E.g., two key equivalences for obtaining temporal normal forms are

$$t\mathcal{U}(p \vee q) \leftrightarrow (t\mathcal{U}p) \vee (t\mathcal{U}q)$$
$$(p \wedge q)\mathcal{U}t \leftrightarrow (p\mathcal{U}t) \wedge (q\mathcal{U}t).$$

In a two-dimensional spatial setting, the first equivalence fails: Fig. 5.19a refutes the implication $\to$. But the other remains a valid principle of monotonicity. The direction $\to$ of the second equivalence is a general monotonicity principle

again. Conversely, we even have a stronger valid law (see Fig. 5.19b for an illustration):

$$p_1 \mathcal{U} q \wedge p_2 \mathcal{U} t \rightarrow (p_1 \wedge p_2) \mathcal{U} (q \vee t).$$

We mention that Aiello, 2002a contains a more sustained analysis of the spatial content of the $\mathbb{R}$ complete Until logic of Burgess, 1984.

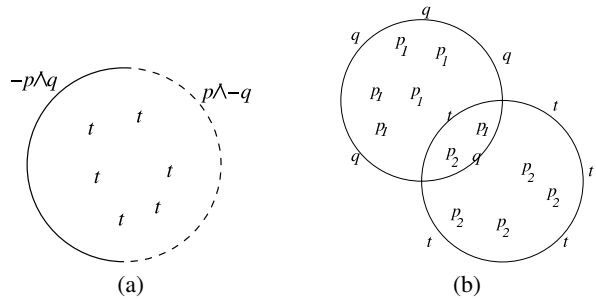


Figure 5.19. Examples of Until models.

3.3.3 Extended spatio-temporal formalisms. Another use for the preceding ideas is in combined logics of *space-time*, treated extensively in Ch. 9. In particular, Shehtman, 1993 axiomatized the complete logic of the rationals in this language, while Gerhardt, 2004 added 'Since'/'Until' using methods of de Jongh and Veltman, 1985 that go back to Burgess, 1979. But axiomatizing the complete logic of the reals remains open. We refer to Bezhanishvili and Kupke, 2006 for some interesting progress in tackling this problem using products of modal logics, while also introducing the 'Since'/'Until' operators for the temporal component. Also worth mentioning are combinations of topological modalities with one for a continuous map giving us the next step in the temporal evolution of some dynamical systems; Ch. 10 explains this approach and provides references to current work, including the recent undecidability results of Konev et al., 2004.

3.4 Topological semantics for epistemic logic

Spatial models can also serve other than geometrical purposes. As we noted in Sec. 1.3, the earliest topological semantics was actually proposed for modelling *intuitionistic logic* based on evidence and knowledge. Nowadays, however, standard relational semantics holds sway in modelling intuitionistic logic or explicit knowledge-based *epistemic logic* in the tradition of Hintikka in philosophy (Hintikka, 1962), Aumann in economics (Binmore, 1994), or Halpern and Parikh in computer science (Fagin et al., 1995; Wooldridge, 2002). Nevertheless, van Benthem and Sarenac, 2004 have shown recently how even

the more technical results obtained in the spatial tradition are illuminating for knowledge once we switch to a topological interpretation. Ch. 8 is also highly relevant in this connection, be it more in the constructive logic tradition.

Our main interest in this chapter is space rather than knowledge. Nevertheless, we sketch some main ideas from van Benthem and Sarenac, 2004, as they involve a further extension of modal languages for space, to also include *fixed-point operators*.

The most-used relational models in epistemic models have reflexive and transitive accessibility relations, and the key semantic clause about an agent's knowledge of a proposition says that $K_i\varphi$ holds at a world x iff φ is true in all worlds y accessible for i from x. For an illustration of how this works cf. Fig. 5.20, where we assume that the 1 and 2 relations are reflexive, transitive, and symmetric.

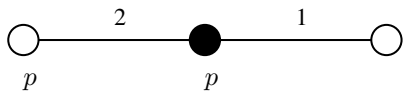


Figure 5.20. In the black central world, 1 does not know if p, while 2 knows that p. Therefore, in the world to the left, 1 knows that 2 knows that p, but 2 does not know if 1 knows p.

Thus, the epistemic knowledge modality is a modal box $\Box_i\varphi$, and the basic logic is that of the spatial interpretation, viz. $\mathbf{S4}$. In an epistemic setting, the spatial modal axioms get a special flavor. E.g., the iteration axiom $\Box_1 p \to \Box_1\Box_1 p$ now expresses 'positive introspection': agents who know something, know that they know it. More precisely, we have $\mathbf{S4}$-axioms for each separate agent, but no further 'mixing axioms' for iterated knowledge of agents such as $\Box_1\Box_2 p \to \Box_2\Box_1 p$. Indeed, the latter implication fails in the above example because in the world on the left, 1 knows that 2 knows that p, but 2 does not know if 1 knows p. Another way of describing the set of valid principles is as the *fusion* of separate logics $\mathbf{S4}$ for each agent. In what follows, we shall mostly work with the two-agent groups $G = \{1, 2\}$.

3.4.1 Group knowledge: agents as relations. A striking discovery in an interactive epistemic setting has been various notions of what may be called *group knowledge*. Two well known examples are the following (Fagin et al., 1995):

1 $E_G\varphi$: every agent in group G knows that φ,

2 $C_G\varphi$: φ is *common knowledge* in the group G.

The latter notion has been proposed in the philosophical, economic, and linguistic literature as a necessary precondition for coordinated behavior between

agents (cf. Lewis, 1969). The usual semantic definition of common knowledge runs as follows:

$$M, x \models C_{1,2}\varphi \text{ iff for all } y \text{ with } x\ (R_1 \cup R_2)^* y \text{ we have } M, y \models \varphi,$$

where $x(R_1 \cup R_2)^* y$ if x is connected to y by a finite sequence of successive steps from either of the two accessibility relations. This is the familiar transitive closure of the union of the relations for both agents. The key valid principles for common knowledge are as follows:

(Equilibrium Axiom) $$C_{1,2}\varphi \leftrightarrow (\varphi \wedge (\Box_1 C_{1,2}\varphi \wedge \Box_2 C_{1,2}\varphi))$$

(Induction Rule) $$\frac{\vdash \varphi \rightarrow (\Box_1(\psi \wedge \varphi) \wedge \Box_2(\psi \wedge \varphi))}{\vdash \varphi \rightarrow C_{1,2}\psi}$$

This logic is known in the literature as $\mathbf{S4}_2^C$. It has been shown to be complete and decidable (Fagin et al., 1995).

A further interesting notion of knowledge for a group of agents is the so-called *implicit knowledge* $D_G\varphi$, which roughly describes what a group would know if its members decided to merge their information:

$$M, x \models D_{1,2}\varphi \text{ iff for all } y \text{ with } x(R_1 \cap R_2)y \text{ we have } M, y \models \varphi,$$

where $R_1 \cap R_2$ is the intersection of the accessibility relations for the separate agents. Unlike universal and common knowledge, this notion is not invariant under modal *bisimulations*. It also involves a new phenomenon of merging information possessed by different agents. The latter topic will return below.

New notions of group knowledge introduce *new agents*. E.g., C_G defines a new kind of $\mathbf{S4}$-agent since $(R_1 \cup R_2)^*$ is again reflexive and transitive. Note that $R_1 \cup R_2$ is not necessarily transitive, so the new 'agent' corresponding to the fact that 'everybody knows' would have different epistemic properties. In particular, it would lack positive introspection as to what it knows. In contrast, the relation $R_1 \cap R_2$ for D_G is again an $\mathbf{S4}$-agent since Horn conditions like reflexivity and transitivity are preserved under intersections of relations. Thus, with two $\mathbf{S4}$-agents $1, 2$, two additional agents supervene, one weaker and one stronger:

$$\begin{array}{ccccc} & & (R_1 \cup R_2)^* & & \\ & \nearrow & & \nwarrow & \\ R_1 & & & & R_2 \\ & \nwarrow & & \nearrow & \\ & & R_1 \cap R_2 & & \end{array}$$

3.4.2 Alternative views of common knowledge. Despite the success of standard epistemic logic, there are still doubts about its expressive power and sensitivity. Notably, the well-known critical paper Barwise, 1988 claimed that a proper analysis of common knowledge must distinguish three different approaches:

1 countably *infinite iteration* of individual knowledge modalities,

2 the *fixed-point view* of common knowledge as 'equilibrium',

3 agents' having a *shared epistemic situation*.

Barwise's distinctions are hard to implement in standard relational semantics. But they make sense in topological semantics—where they suggest interesting language extensions. Here is some technical groundwork.

The Equilibrium Axiom for the operator $C_G\varphi$ describes it as a fixed-point of an epistemic operator $\lambda X.\varphi \wedge \Box_1 X \wedge \Box_2 X$. In conjunction with the Induction Rule, it may even be seen to be the *greatest fixed-point* definable in the standard modal μ-*calculus* as:

$$C_G\varphi := \nu p.\varphi \wedge \Box_1 p \wedge \Box_2 p.$$

The greatest fixed-point is computed as the first stabilization stage of a descending approximation sequence for a *monotonic* set function through the ordinals. We write $[|\varphi|]$ for the truth set of φ in the relevant model where evaluation takes place:

$C^0_{1,2}\varphi := [|\varphi|]$,
$C^{\kappa+1}_{1,2}\varphi := [|\varphi \wedge \Box_1(C^\kappa_{1,2}\varphi) \wedge \Box_2(C^\kappa_{1,2}\varphi)|]$,
$C^\lambda_{1,2}\varphi := [|\bigwedge_{\kappa<\lambda} C^\kappa_{1,2}\varphi|]$, for λ a limit ordinal.

Finally, we let $C_{1,2}\varphi := C^\kappa_{1,2}\varphi$ where κ is the least ordinal for which the approximation procedure halts; that is, $C^{\kappa+1}_{1,2}\varphi = C^\kappa_{1,2}\varphi$. In general, reaching this stopping point may take any number of ordinal stages. E.g., the least-fixed-point formula $\mu p.\Box p$ computing the 'well-founded part' of the binary

accessibility relation may stabilize only at the cardinality of the model. But in certain cases we can do much better, as the following well-known fact shows:

FACT 5.84 *In every relational epistemic model, the approximation procedure for the common knowledge modality stabilizes at $\kappa \leq \omega$.*

This result shows that Barwise's fixed-point and countable-iteration views of common knowledge coincide in relational models. More precisely, $\nu p.\varphi \wedge \Box_1 p \wedge \Box_2 p$ is equivalent to

$$K_{1,2}\varphi := \varphi \wedge \Box_1\varphi \wedge \Box_2\varphi \wedge \Box_1\Box_2\varphi \wedge \ldots \quad .$$

The simple stabilization behavior at ω is most easily understood by observing that the knowledge modalities $\Box_i$ distribute over any infinite conjunction. Therefore, $\Box_i(\bigwedge_{n<\omega} C^n_{1,2}\varphi)$ is simply $\bigwedge_{n<\omega} \Box_i C^n_{1,2}\varphi$ which is equivalent to $\bigwedge_{n<\omega} C^n_{1,2}\varphi$. More generally, stabilization for the formula $\nu p.\varphi(p)$ is guaranteed by stage ω in any model in case the syntax defining the monotone approximation operator is constrained to a 'universal-conjunctive' format (van Benthem, 1996).

3.4.3 Topological models for epistemic logic with fixed-points. The language of epistemic logic can be interpreted just as well in topological models, although the presence of many agents calls for an indexed *family of topologies* on the base set of worlds, representing their individual information structures. All the notions such as bisimulation, axiomatic systems, and the product constructions of Sec. 3.2 also make sense epistemically. But these now acquire a special flavor—putting together topological models into one product space amounts to *merging information spaces* for different agents. The earlier horizontal and vertical topologies on the products encode the agents' original individual spaces. Our earlier result that the modal logic of the product construction is the fusion $\mathbf{S4} \oplus \mathbf{S4}$ then says epistemically that we have really defined a good 'conservative merge' without side-effects.

Further topologies on the product space encode further emergent group-oriented information structures. The earlier definitions of common knowledge still make sense in topological models. As before, the countably infinite iteration of all finite sequences of alternating knowledge modalities for the individual agents $1, 2$ is

$$K_{1,2}\varphi := \bigwedge_{n<\omega} K^n_{1,2}\varphi,$$

with $K^n_{1,2}\varphi$, defined inductively as follows:

$$K^0_{1,2}\varphi := \varphi$$
$$K^{n+1}_{1,2}\varphi := \Box_1(K^n_{1,2}\varphi) \wedge \Box_2(K^n_{1,2}\varphi).$$

The same is true for the fixed-point definition

$$C_{1,2}\varphi := \nu p.\varphi \wedge \Box_1 p \wedge \Box_2 p.$$

However, the definitions of common knowledge by fixed-points and by countably infinite iteration will now diverge, because one can show that given an interpretation of p, the interpretation of $K_{1,2}p$ does not always define a horizontally and vertically open set in the product model. Since the fixed-point version of $C_{1,2}p$ is always horizontally and vertically open, it follows that the two are not the same. We refer to van Benthem and Sarenac, 2004 for the details.

Returning to an earlier topic, we can now view *product spaces* as introducing new 'collective agents' via new topologies. In particular, common knowledge as a greatest fixed-point corresponds to the *intersection* $\tau_1 \cap \tau_2$ of the horizontal and vertical topologies on the product space. On the other hand, the topological meaning of the implicit group knowledge D_G is the *join* $\tau_1 \vee \tau_2$ of the horizontal and vertical topologies. Its basis is the pairwise intersection of horizontal and vertical opens. The latter topology need not always be of great interest. For instance, $\tau_1 \vee \tau_2$ is discrete on $\mathbb{Q} \times \mathbb{Q}$. From an informational perspective, this means that merging the information that we get about points in the horizontal and vertical directions fixes their position uniquely. All this again yields an inclusion diagram:

$$\begin{tikzcd}
 & \tau_1 \vee \tau_2 & \\
\tau_1 \arrow[ur] & & \tau_2 \arrow[ul] \\
 & \tau_1 \cap \tau_2 \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Returning to the three distinctions made in Barwise, 1988, what about the third view of having a 'shared situation'? One good candidate for it would be the standard product topology τ. The agent corresponding to this new group concept τ only accepts very strong collective evidence for any proposition. And we know the complete logic of adding this agent from the joint axiomatization of horizontal, vertical, and standard product topologies from Sec. 3.2.3.

4. Modal logic and geometry

This chapter has been mainly concerned with modal aspects of topology. But many mathematical theories of space exist beyond topology, such as affine and metric geometry, or newer theories like mathematical morphology, more

within linear algebra. And indeed, modal structures emerge in all of them. Our final two sections provide a brief account of this, if only to put the modal topological approach in a broader perspective. Sec. 4 is about geometry proper, Sec. 5 will deal with vector spaces. For further details and alternative logical approaches, we refer to the chapters in this book on modal and first-order theories of geometry (Ch. 2; Ch. 7; Ch. 14).

We start by recalling that affine geometry is given by the following three axioms involving points, lines, and an incidence relation (Blumenthal, 1961; Goldblatt, 1987; and Ch. 7):

A1 Any two distinct points lie on exactly one line.

A2 There exist at least three non-collinear points.

A3 Given a point a and a line L, there is exactly one line M that passes through a and is parallel to L.

Affine spaces have a strong modal flavor (Balbiani et al., 1997; Balbiani, 1998; Venema, 1999; Stebletsova, 2000). Approaches include two-sorted versions with matching bimodal operators, and merging points and lines into one sort of pairs $\langle$point, line$\rangle$ equipped with two incidence relations. By contrast, the classical approach to affine structure is Tarski, 1959, which contains a complete first-order axiomatization of elementary geometry in terms of a ternary betweenness predicate $\beta(xyz)$, as well as quaternary equidistance $\delta(xyzu)$, interpreted as x is as distant from y as z is from u. Yet, Tarski's beautiful decidable axiomatization still leaves things to be desired. First, the system has high complexity, viz. exponential space (Ben-Or et al., 1986). And from an expressive viewpoint, the axioms mix betweenness and equidistance, whereas one would like to understand affine and metric structure separately. A complete axiomatization of pure affine first-order geometry was given in Szczerba and Tarski, 1965.

We now turn to the modal view. There will be a 'style break' here as compared to our account in this chapter so far. We presented topological models as a generalization of relational semantics, where the topology need not be generated by any ordering relation on the space. In visual terms, this reflects the intuitively 'malleable' nature of open sets. However, moving from topology to geometry, we encounter basic relations such as betweenness and equidistance. And these lead to modal structures of a more classical relational kind, be it with *ternary* rather than the usual binary relations. This difference should not be a problem, but rather an asset. Indeed, topological and relational views live together harmoniously, say in a combined modal language for topology plus betweenness.

4.1 Affine geometry in modal logic

4.1.1 Basic modal language and affine transformations. We start by defining a binary betweenness modality $<B>$:

$$M, x \models <B>(\varphi, \psi) \quad \text{iff} \quad \exists y, z : \beta(yxz) \text{ and } M, y \models \varphi \text{ and } M, z \models \psi.$$

Our language is a propositional language enriched with the betweenness modal operator $<B>$. Models for this language are triples $\langle X, \beta, \nu \rangle$, where X is a nonempty set, β is a ternary betweenness relation on X, and ν is a valuation function. In this setting, familiar modal notions acquire sometimes surprising new flavors. For instance, we get a new sort of geometrical model transformation appropriate to the modal view of affine structure. *Affine bisimulations* relate points verifying the same proposition letters, while maintaining the betweenness relation:

DEFINITION 5.85 (AFFINE BISIMULATION) Given two affine models $\langle X, \beta, \nu \rangle$ and $\langle X', \beta', \nu' \rangle$, with x, y, z ranging over X and x', y', z' over X', an *affine bisimulation* is a nonempty relation $B \subseteq X \times X'$ such that if xBx' then:

1 x and x' satisfy the same proposition letters

2 (forth condition): $\beta(yxz) \rightarrow \exists y'z' : \beta'(y'x'z')$ and yBy' and zBz'

3 (back condition): $\beta'(y'x'z') \rightarrow \exists yz : \beta(yxz)$ and yBy' and zBz'.

In Goldblatt, 1987 isomorphisms are considered the only interesting maps across affine models. But in fact, just as with topological bisimulations versus homeomorphisms (Theorem 5.6), affine bisimulations are interesting coarser ways of comparing spatial situations. In the true modal spirit they only consider behavior of points inside local line environments. Fig. 5.21 shows a case of non-isomorphic yet bisimilar triangles with atomic properties indicated. This affine bisimulation can be regarded as a sort of 'modal contraction' to the smallest model with the same structure.

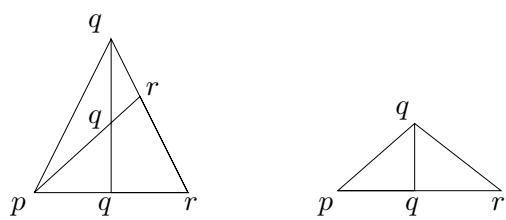


Figure 5.21. Affine bisimilar models.

By contrast, the models in Fig. 5.22 are not bisimilar: affine bisimulations preserve truth of modal formulas in an obvious way, but $q \wedge <B>(r, r)$ holds

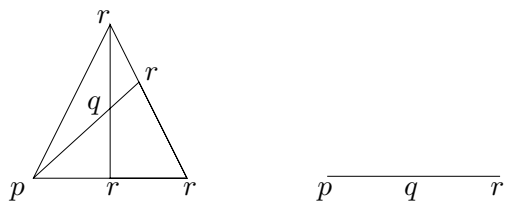


Figure 5.22. Non affine bisimilar reduction.

at the q point of the left model and nowhere on the right. This is also 'modal logic of space', but clearly in a new vein!

Incidentally, there *is* a smaller affine bisimulation contraction for the left-hand triangle in Fig. 5.22. But the resulting model is not 'planar': it cannot be represented in two-dimensional Euclidean space. Now consider a new valuation shown in Fig. 5.23. In this case there does *not* exist a bisimilar contraction: every point of the triangle is distinguishable by a formula which is not true on any other point, see Fig. 5.24.

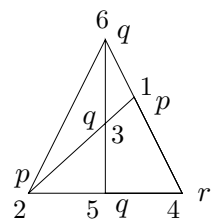


Figure 5.23. An irreducible affine model.

Point	Formula
1	$\varphi_1 = \quad p \wedge <B>(q, r)$
2	$\varphi_2 = \quad p \wedge \neg\varphi_1$
3	$\varphi_3 = \quad q \wedge <B>(\varphi_1, \varphi_2)$
4	$\varphi_4 = \quad r$
5	$\varphi_5 = \quad q \wedge <B>(\varphi_2, \varphi_4)$
6	$\varphi_6 = \quad q \wedge \neg\varphi_3 \wedge \neg\varphi_5$

Figure 5.24. Formulas true at points of the model in Fig. 5.23.

4.1.2 Modal logics of betweenness. The preceding language has a minimal logic as usual, which does not yet have much geometric content. Its

key axioms are two distribution laws:

$$<B>(p \vee q, r) \leftrightarrow <B>(p, r) \vee <B>(q, r)$$
$$<B>(p, q \vee r) \leftrightarrow <B>(p, q) \vee <B>(p, r).$$

This minimal logic has all the usual modal properties, including decidability. Further axioms would express basic universal frame conditions such as betweenness being symmetric at end-points and all points lying 'in between themselves':

$$<B>(p, q) \rightarrow <B>(q, p)$$
$$p \rightarrow <B>(p, p).$$

These are simple modal *frame correspondences*. A more interesting example was already mentioned in Sec. 1.5, involving an existential affine axiom. Consider *associativity* of the betweenness modality:

$$<B>(p, <B>(q, r)) \rightarrow <B>(<B>(p, q), r).$$

FACT 5.86 *Modal Associativity corresponds to Pasch's Axiom.*

Proof We spell out the simple correspondence argument to show how easy matches can be between modal axioms and geometric laws. Consider Pasch's Axiom (Sec. 1.5). Suppose that

$$\forall txyzu(\beta(xtu) \;\&\; \beta(yuz) \rightarrow \exists v : \beta(xvy) \;\&\; \beta(vtz))$$

holds in a frame. Assume that a point t satisfies $<B>(p, <B>(q, r))$. Then there exist points x, u with $\beta(xtu)$ such that $x \models p$ and $u \models <B>(q, r)$. Therefore, there also exist points y, z with $\beta(yuz)$ such that $y \models q$ and $z \models r$. Now by Pasch's Axiom, there must be a point v with $\beta(xvy)$ and $\beta(vtz)$. Thus, $v \models <B>(p, q)$ and hence $t \models <B>(<B>(p, q), r)$.

Conversely, assume that $\beta(xtu)$ and $\beta(yuz)$. Define a valuation on the space by setting $\nu(p) = \{x\}$, $\nu(q) = \{y\}$, and $\nu(r) = \{z\}$. Thus, $u \models <B>(q, r)$ and

$$t \models <B>(p, <B>(q, r)).$$

Then by the validity of Modal Associativity,

$$t \models <B>(<B>(p, q), r).$$

Therefore, there must be points v, w with $\beta(vtw)$ such that $v \models <B>(p, q)$ and $w \models r$. By the definition of ν, the latter implies $w = z$, and the former that $\beta(xvy)$. So indeed v is the required point. QED

All these correspondences may even be *computed automatically* as they have 'Sahlqvist form' (cf. Blackburn et al., 2001 for more general theory). Thus,

modal axioms correspond to significant geomterical axioms, reflecting insights from the usual arithmetizations of geometry. Deeper examples may be found in Stebletsova, 2000 and Stebletsova and Venema, 2001, who analyze modal structures in projective geometry, including Pappus' Theorem.

Complete affine modal logics of special models may also be axiomatized, though only few examples have been dealt with so far. At least for the real line $I\!R$, the task is easy as one can take advantage of the binary ordering $\leq$, defining

$$M, x \models <B>(\varphi, \psi) \quad \text{iff} \quad \exists y, z : \; M, y \models \varphi \text{ and } M, z \models \psi \text{ and } y \leq x \leq z.$$

Using this, we can define temporal operators Future and Past (both including the present). Conversely, these two unary operators define $<B>$ on $I\!R$:

$$<B>(\varphi, \psi) \leftrightarrow (P\varphi \wedge F\psi).$$

Thus, a complete and decidable axiomatization for our $<B>$-language can be found using the well-known tense logic of future and past on $I\!R$ (Segerberg, 1970). We refer to Ch. 7 for further details.

4.1.3 Logics of convexity. An interesting and rich special case of affine structure is the *convex closure* of a set, consisting of all points lying on a segment whose end-points are in the set. Convexity is important in many fields from computational geometry (Preparata and Shamos, 1985) to cognitive science (Gärdenfors, 2000). We can capture convexity modally by frames of points with the betweenness relation:

(5.2) $$M, x \models C\varphi \text{ iff } \exists y, z : M, y \models \varphi \text{ and } M, z \models \varphi \text{ and } x \text{ lies in between } y \text{ and } z.$$

This is a *one-step convexity* operator whose countable iteration yields the standard convex closure, cf. Fig. 5.6. The corresponding binary modality $C\varphi$ is defined as follows:

$$\exists yz : \; \beta(yxz) \text{ and } \varphi(y) \text{ and } \varphi(z).$$

Basic axioms are different here from the preceding subsection. In particular, distributivity fails. The one-step convex closure of a set of two distinct points is the whole interval in between, while the union of their separate one-step closures is just these points themselves. Thus, only monotonicity remains as a valid reasoning principle. Another principle which may fail is the idempotence of the convexity modality:

$$CC\varphi \leftrightarrow C\varphi.$$

Iterating $C\varphi$ can lead to new sets, witness Fig. 5.6. Even so, the non-idempotence is of interest, as it helps distinguish dimensions. For instance, $CC\varphi \leftrightarrow C\varphi$ holds in $I\!R$, but not in $I\!R^2$.

One may now think that the stages $C^{n+1}\varphi \leftrightarrow C^n\varphi$ determine the dimension of the spaces $\mathbb{R}^n$ for all n. But here is a surprise.

THEOREM 5.87 (AIELLO, 2002A) *$CCC\varphi \leftrightarrow CC\varphi$ holds in $\mathbb{R}^3$.*

Nevertheless, convexity does provide dimension principles after all. Here is an old result from Helly, 1923:

THEOREM 5.88 If $K_1, K_2, \ldots, K_m$ are convex sets in the n-dimensional Euclidean space $\mathbb{R}^n$ with $m > n+1$, and if for every choice of $n+1$ sets K_i there exists a point that belongs to all the chosen sets, then there exists a point that belongs to all the sets $K_1, K_2, \ldots, K_m$.

Our modal language formalizes this theorem as follows:

$$\bigwedge_{f:\{1,\ldots,n+1\}\to\{1,\ldots,m\}} E(\bigwedge_{i=1}^{n+1}(C^n\varphi_{f(i)}) \to E(\bigwedge_{i=1}^{m} C^n\varphi_i),$$

where E is the existential modality defined in terms of betweenness:

$$E\varphi \text{ iff } <B>(\varphi, \top).$$

Interestingly, Helly's results have been revived recently in Leitgeb, 2005 in a new reconstruction of Carnap's geometrical constructions in Carnap, 1998. Thus again, modal languages capture significant geometrical facts.

4.1.4 First-order affine geometry. As usual, the above modal language is a fragment of a first-order language under the standard translation. The relevant first-order language is not quite that of Tarski's elementary geometry, however, as we also get unary predicate letters denoting regions. That is, in terms of validity, we are rather in *monadic second-order logic*. As in our discussion of topology, the affine first-order or monadic second-order language of regions is a natural limit toward which affine modal languages can strive via various logical extensions. From a geometrical viewpoint, one might also hope that 'layering' the usual language in this modal way will bring to light interesting new geometrical facts. For much more information on current first-order theories of geometry we refer to Ch. 2, Ch. 7.

Another major feature of standard geometry is the *equal status of points and lines*. This would suggest a reorganization of the modal logic to a *two-sorted* one (cf. Marx and Venema, 1997) stating properties of both points and segments viewed as independent semantic objects. One can think of this as a way of lowering the second-order complexity as the relevant subsets have now become first-order objects in their own right (cf. van Benthem, 1999).

There are also other analogies with existing modal systems, such as *arrow logic* (van Benthem, 1996; Venema, 1996), where 'arrows'—standing for directed transitions between points (or vectors)—become semantic objects in their own right, in addition to the points themselves. Arrow logic will be explained in some more detail in our account of mathematical morphology and linear algebra. The two-sorted move seems very geometrical in spirit, and it would also reflect duality principles of the sort that led from affine to projective geometry.

4.2 Metric geometry in modal logic

The next level in geometry beyond affine point-line patterns is *metric* structure. We show how modal languages can describe a notion of comparative distance.

4.2.1 Structures for relative nearness. Relative nearness was introduced in van Benthem, 1983b as a natural primitive for 'orientation' which has spatial discourse in natural language:

$N(x, y, z)$ iff y is closer to x than z is, i.e., $d(x, y) < d(x, z)$
where $d(x, y)$ is the distance function (see Fig. 5.25).

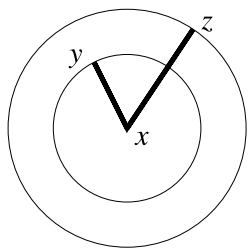


Figure 5.25. From point x, point y is closer than point z.

The function d can be a spatial metric, cognitive visual closeness, or even a utility function. Randell et al., 2001 develop the first-order logic of comparative nearness for the purpose of robot navigation, extending the calculus of regions RCC. In such a setting, relative nearness is a powerful primitive, as it defines equidistance:

$$Eqd(x, y, z) : \neg N(x, y, z) \text{ and } \neg N(x, z, y).$$

Affine betweenness is also definable in terms of N, at least in $\mathbb{R}^n$: cf. Sec. 4.2.2. Finally, even the identity of points $x = y$ is expressible:

$$x = y \text{ iff } \neg N(x, x, y).$$

The further analysis of this structure can proceed as in the affine case. But there are also surprising open questions. E.g., the *universal first-order theory* of relative nearness for Euclidean spaces has still not been axiomatized.

4.2.2 Modal logic of nearness. In line with the main concern of this chapter, consider this obvious modal operator accessing ternary nearness N:

$$M, x \models <N>(\varphi, \psi) \text{ iff } \exists y, z : \ M, y \models \varphi \text{ and } M, z \models \psi \text{ and } N(x, y, z).$$

The universal dual of $<N>$ is also interesting in its spatial behavior:

$$M, x \models [N](\varphi, \psi) \text{ iff } \forall y, z \ (N(x, y, z) \text{ and } M, y \models \varphi \rightarrow M, z \models \neg\psi).$$

Dropping the negation, one gets the following appealing notion:

> If any point y around the current point x satisfies φ, then all points z further out must satisfy ψ.

The basic modal logic of nearness again has distribution laws:

$$<N>(p \vee q, r) \leftrightarrow <N>(p, r) \vee <N>(q, r)$$
$$<N>(p, q \vee r) \leftrightarrow <N>(p, q) \vee <N>(p, r).$$

On top of this, natural universal frame constraints return as special axioms. Here are two examples:

(transitivity)
$$<N>(p, q) \wedge \neg<N>(p, p) \wedge \neg<N>(q, q) \wedge <N>(q, r) \rightarrow <N>(p, r)$$

(connectedness)
$$<N>(p, q) \wedge \neg<N>(p, p) \wedge \neg<N>(q, q) \wedge Er \rightarrow <N>(p, r) \vee <N>(r, q).$$

Modal logics of nearness on special structures may include further constraints computable by correspondence techniques. Here is a useful observation covering many cases. Our language can define that φ holds in just one unique point:

$$E!\,\varphi \text{ iff } E(\varphi \wedge \neg<N>(\varphi, \varphi)).$$

Now a straightforward proof, known from extended modal logics with a difference modality (de Rijke, 1993), establishes the following:

PROPOSITION 5.89 *Every universal first-order property of N is modally definable.*

Sheremet et al., 2006 provide a sophisticated analysis of modal languages for comparative nearness, including boundaries between decidable and undecidable systems.

4.2.3 First-order theory of nearness. As in the affine case, the backdrop to our modal analysis is the first-order theory of relative nearness (cf. Aiello and van Benthem, 2002b).

FACT 5.90 *The single primitive of comparative nearness defines the two primitives of Tarski's Elementary Geometry in first order logic.*

Proof The following defines betweenness (see Fig. 5.26):

$$\beta(yxz) \text{ iff } \neg\exists x' : N(y, x', x) \text{ and } N(z, x', x).$$

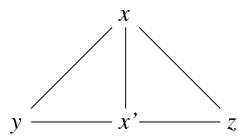


Figure 5.26. Defining betweenness via nearness.

This allows us to define parallel segments in the usual way as having no intersection points on their generated lines:

$$\begin{aligned} xx'||yy' \leftrightarrow & \neg\exists c : \beta(xx'c) \wedge \beta(yy'c) \wedge \\ & \neg\exists c' : \beta(c'xx') \wedge \beta(c'yy') \wedge \\ & \neg\exists c'' : \beta(xc''x') \wedge \beta(yc''y'). \end{aligned}$$

Then one defines equal segment length by

$$\delta(x, y, z, u) \text{ iff } \exists y' : \ xu||yy' \text{ and } xy||uy' \text{ and } \neg N(u, z, y') \text{ and } \neg N(u, y'z)$$

(see Fig. 5.27).

QED

There are many other systems of first-order geometry with similar richness. For instance, see the axiomatization of constructive geometry in von Plato, 1995. Having established this connection with the first-order realm, we pass the torch to the relevant chapters in this book: Ch. 2, Ch. 7.

5. Modal logic and linear algebra

Classical geometry is not the last word in mathematical theories of space. Connections between *linear algebra* and spatial representation are well-known from a major qualitative visual theory, viz. *mathematical morphology* (Matheron, 1967; Serra, 1982). We provide a brief treatment along the lines of Aiello

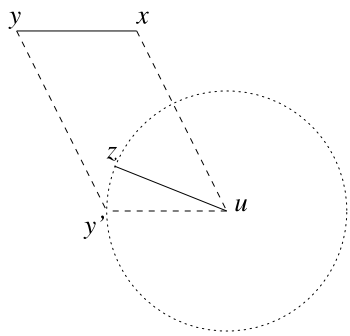


Figure 5.27. Equidistance in terms of nearness.

and van Benthem, 2002a and van Benthem, 2000b, to show that modal structures play a natural role here as well. The flavor of modal logic of linear algebra is different from that of topology or geometry, but similar themes emerge all the same. A different connection between mathematical morphology and modal logic can be found in Bloch, 2000. See Ch. 14 of this book for a more comprehensive treatment.

5.1 Mathematical morphology

We first set the scene. In line with our spatial interests, we consider vector spaces $\mathbb{R}^n$. 'Images' are regions consisting of sets of vectors. Mathematical morphology provides four basic ways of combining or simplifying images, viz. *dilation*, *erosion*, *opening* and *closing*. These are illustrated in Fig. 5.28.

Intuitively, dilation adds regions together while, e.g., erosion is a way of removing 'measuring idiosyncrasies' from a region A by using region B as a kind of boundary smoothener. (If B is a circle, one can think of it as rolling tightly along the inside of A's boundary, leaving only a smoother interior version of A.) More formally, dilation or *Minkowski addition* $\oplus$ is a sum operation on sets of vectors:

$$A \oplus B = \{a + b \ : \ a \in A \text{ and } b \in B\} \quad \text{dilation}$$

This is naturally accompanied by

$$A \ominus B = \{a \ : \ a + b \in A \text{ for all } b \in B\} \quad \text{erosion}$$

Openings and closings are combinations of dilations and erosions:

the structural *opening* of A by B	$(A \ominus B) \oplus B$
the structural *closing* of A by B	$(A \oplus B) \ominus B$

Mathematical morphology also employs the usual boolean operations on regions: intersection, union, and complement. In addition to topology and

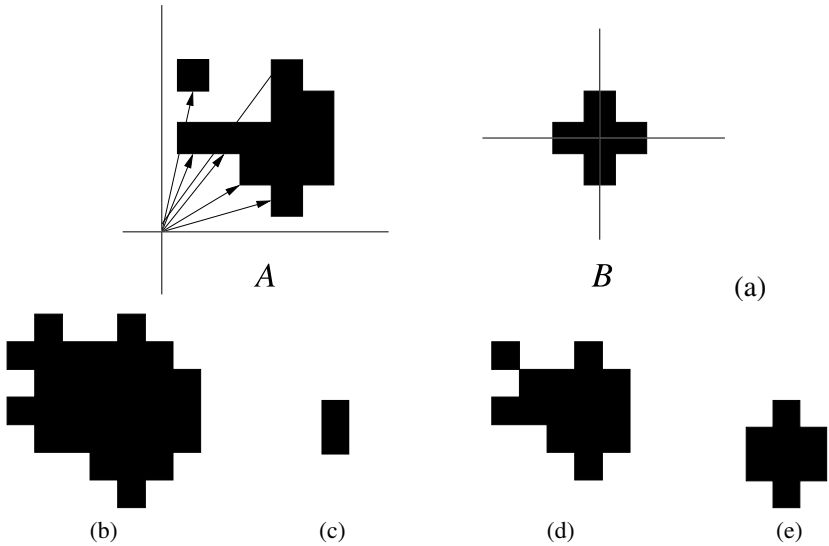


Figure 5.28. (a) Regions A and B of the vector space $\mathbb{R}^2$; (b) dilating A by B; (c) eroding A by B; (d) closing A by B; (e) opening A by B.

geometry, this is yet another mathematical theory of $\mathbb{R}^n$, this time focusing on their vector structure. Evidently, the above operations are only a small sub-calculus, chosen for its computational utility and expressive perspicuity.

5.2 First step: links with linear logic

Mathematical morphology is not normally thought of in terms of logic, but on reflection, it does have a logical flavor. The Minkowski operations behave like notions of *propositional logic* in some procedural mode. Dilation is like a logical conjunction $\oplus$, and erosion like an implication $\rightarrow$, as seems clear from their definitions ('combining an A and a B', and 'if you give me a B, I will give you an A'). The two are related by the following *residuation law*:

$$A \oplus B \subseteq C \text{ iff } A \subseteq B \rightarrow C$$

which is also typical for conjunction and implication. Nevertheless, there are also some differences. For instance, $A \oplus A$ is not in general equal to A.

A first logical calculus for these operations (not yet 'modal', but see Sec. 5.3) is known as *multiplicative linear logic* in computer science and as the *Lambek calculus with permutation* in linguistics (Troelstra, 1992; Kurtonina, 1995). The calculus derives 'sequents' of the form $A_1, \ldots, A_k \Rightarrow B$ where

each expression A, B in the current setting stands for a region, and the intended interpretation—in our case—says that

> The sum of the A's is included in the region denoted by B.

Here are the derivation rules, starting from basic axioms $A \Rightarrow A$:

(product rules) $$\frac{X \Rightarrow A \qquad Y \Rightarrow B}{X, Y \Rightarrow A \oplus B} \qquad\qquad \frac{X, A, B \Rightarrow C}{X, A \oplus B \Rightarrow C}$$

(arrow rules) $$\frac{A, X \Rightarrow B}{X \Rightarrow A \to B} \qquad\qquad \frac{X \Rightarrow A \qquad B, Y \Rightarrow C}{X, A \to B, Y \Rightarrow C}$$

(structural rules) $$\frac{X \Rightarrow A}{\pi[X] \Rightarrow A}\,\text{permutation} \qquad\qquad \frac{X \Rightarrow A \qquad A, Y \Rightarrow B}{X, Y \Rightarrow B}\,\text{cut}$$

Derivable sequents typically include:

('function application') $A, A \to B \Rightarrow B$

('function composition') $A \to B, B \to C \Rightarrow A \to C$

Other key examples are two 'currying' laws, provable using $\oplus$ rules:

$$(A \oplus B) \to C \Rightarrow (A \to (B \to C))$$
$$(A \to (B \to C)) \Rightarrow (A \oplus B) \to C.$$

The major combinatorial properties of this calculus **LL** are known, including proof-theoretic cut elimination theorems, and *decidability* of derivability in *NP* time. Moreover, there are several formal semantics underpinning this calculus (algebraic, game-theoretic, category-theoretic, and possible worlds-style; see van Benthem, 1991a. On top of all this, mathematical morphology provides a new model for linear logic!

FACT 5.91 (AIELLO AND VAN BENTHEM, 2002B) *Every space $\mathbb{R}^n$ with the Minkowski operations is a model for all* **LL***-provable sequents.*

This soundness theorem shows that every sequent derivable in **LL** must be a valid principle of mathematical morphology. The converse seems an open question of independent interest.

Further laws of mathematical morphology 'mix' pure Minkowski operations $\oplus, \to$ with standard boolean ones. E.g. they include the fact that $(A \cup B) \to C$ is the same as $(A \to C) \cap (B \to C)$. This requires adding boolean operations to **LL**:

$$\frac{X, A \Rightarrow B}{X, A \cap C \Rightarrow B} \qquad \frac{X, A \Rightarrow B}{X, C \cap A \Rightarrow B} \qquad \frac{X \Rightarrow A \qquad X \Rightarrow B}{X \Rightarrow A \cap B}$$

$$\frac{X \Rightarrow A}{X \Rightarrow A \cup B} \qquad \frac{X \Rightarrow A}{X \Rightarrow B \cup A} \qquad \frac{X, A \Rightarrow B \qquad X, C \Rightarrow B}{X, A \cup C \Rightarrow B}$$

5.3 Arrow logic and hybrid modalities

But what about modal logic, the main theme of this chapter? We will see that there is a natural connection after all.

The basic players in the above algebra of regions in a vector space are surely the vectors themselves. For instance, Fig. 5.28.a represents the region A as a set of vectors departing from the origin. Vectors come with some natural operations such as binary addition or unary inverse—witness the usual definition of a vector space. A vector v in our particular spaces may be viewed as an ordered pair of points (o, e), with o the origin and e the end point. Pictorially, this is an arrow from o to e. Now this provides a point of entry into one more area of modal logic.

Arrow logic was developed in the early 1990's as a modal theory of transitions or arrows structured by various relations. In particular, there is a binary modality for *composition* of arrows and a unary one for *converse*. The motivation comes from dynamic logics, treating transitions as objects in their own right, and from relational algebra, making pairs of points into separate objects. This allows for greater expressive power than the usual algebraic systems, while also lowering complexity of the core logics (Blackburn et al., 2001; van Benthem, 1996). For instance, the pair-interpretation has arrows as pairs of points (a_o, a_e), and then defines these semantic relations:

composition $C(a_o, a_e)(b_o, b_e)(c_o, c_e)$ iff $a_o = c_o$, $a_e = b_o$, and $b_e = c_e$,

inverse $R(a_o, a_e)(b_o, b_e)$ iff $a_o = b_e$ and $a_e = b_o$,

identity $I(a_o, a_e)$ iff $a_o = a_e$.

An abstract model is a set of arrows as primitive objects, with three relations as above, and a valuation function as usual:

DEFINITION 5.92 (ARROW MODEL) An *arrow model* is a tuple $M = \langle W, C, R, I, \nu \rangle$ such that $C \subseteq W \times W \times W$, $R \subseteq W \times W$, $I \subseteq W$, and $\nu : P \to \mathcal{P}(W)$.

Such models have a wide variety of interpretations, from linguistic syntax to category theory (Venema, 1996). And they can be made even more useful by having a two-sorted version with both *points and arrows* as primitive objects.

Now introduce a standard modal language with proposition letters, the identity element 0, monadic operators $\neg, -$, and a dyadic operator $\oplus$. The truth definition then has the following obvious clauses:

$$
\begin{array}{lll}
M, x \models p & \text{iff} & x \in \nu(p) \\
M, x \models 0 & \text{iff} & Ix \\
M, x \models -\varphi & \text{iff} & \exists y : \ Rxy \text{ and } M, y \models \varphi \\
M, x \models \neg\varphi & \text{iff} & \text{not } M, x \models \varphi \\
M, x \models \varphi \vee \psi & \text{iff} & M, x \models \varphi \text{ or } M, x \models \psi \\
M, x \models A \oplus B & \text{iff} & \exists y \exists z : \ Cxyz \text{ and } M, y \models A \text{ and } M, z \models B \\
M, x \models A \ominus B & \text{iff} & \forall yz \ (Cyxz \text{ and } M, z \models A \rightarrow M, y \models B).
\end{array}
$$

This language leads to a decidable basic system of arrow logic:

$$(p \vee q) \oplus r \leftrightarrow (p \oplus r) \vee (q \oplus r) \tag{5.3}$$

$$p \oplus (q \vee r) \leftrightarrow (p \oplus q) \vee (p \oplus r) \tag{5.4}$$

$$-(p \vee q) \leftrightarrow (-p \wedge -q) \tag{5.5}$$

$$p \wedge (q \oplus r) \rightarrow q \oplus (r \wedge (-q \oplus p)). \tag{5.6}$$

This connects relational algebra, linear logic, and categorial grammar (cf. Kurtonina, 1995).

Important to us here, however, is the obvious connection with *vector spaces*. Think of the abstract composition $Cxyz$ as vector addition $x = y + z$, of converse Rxy as vector inverse $x = -y$, and of identity elements Ix as the null-vector $x = 0$. Most modal topics now make immediate sense in linear algebra or mathematical morphology. E.g., arrow models support a natural notion of *bisimulation*, which will now compare vector spaces in coarser ways than their usual linear transformations.

Next, there is valid reasoning. The laws of basic Arrow Logic represent obvious vector laws. E.g., to see the validity of (5.6), note that if a vector a is the composition of b and c, then c can also be written as the composition $-b \oplus a$.

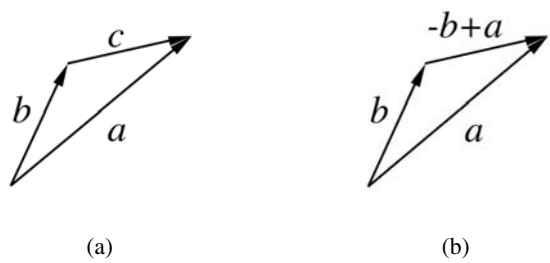


Figure 5.29. Triangle axiom for arrow composition.

On top of the base system, special arrow logics have been axiomatized with a number of additional frame conditions (Marx, 1995; Mikulas, 1995) and a

resolution calculus has been introduced for the purpose of knowledge representation and spatial reasoning (Aiello and Ottens, 2007). In particular, the vector space interpretation makes composition *commutative* and *associative*, validating two further axioms:

(commutativity) $$p \oplus q \leftrightarrow q \oplus p$$

(associativity) $$p \oplus (q \oplus r) \leftrightarrow (p \oplus q) \oplus r$$

The key fact about composition is now the vector law

$$a = b + c \quad \text{iff} \quad c = a - b$$

which derives the triangle inequality (see Fig. 5.29).

Again the soundness of arrow logic is clear, and we can freely derive old and new laws of vector algebra in the above calculus. But our connection between arrow logic and mathematical morphology also raises an intriguing open problem: *What is the complete axiomatization of arrow logic over the standard vector spaces* $\mathbb{R}^n$*?*

6. Conclusions

The work surveyed in this chapter suggests that modal logic is a natural medium for analyzing reasoning about spatial patterns. In particular, modal research into topological structure is showing signs of becoming a recognized topic with a fast-growing agenda, as shown in Sec. 2, 3. And this interface has interesting repercussions in both directions. Modal logic acquires new spatially inspired models and new questions about them, while topology acquires new modally inspired notions, such as bisimulation or complete calculi for language fragments. We have also shown briefly (in Sec. 4, 5) how the modal perspective can be extended to deal with further geometrical structure: affine, metric, and even vector-based, more in the spirit of classical graph-like relational models. Again, benefits of taking such a stance can flow in surprising ways. We have indicated, e.g., how spatial models can also inject new ideas into areas such as dynamic or epistemic logic. In our view, all this goes to show that, just as temporal logic has been for so long, spatial logic might become a focus for modal logic.

Acknowledgement

We are grateful to David Gabelaia for his careful proofreading and comments.

References

Abashidze, M. (1987). *Algebraic Analysis of the Gödel-Löb Modal System*. PhD thesis, Tbilisi State University. In Russian.

Abashidze, M. and Esakia, L. (1987). Cantor's scattered spaces and the provability logic. In *Baku International Topological Conference. Volume of Abstracts. Part I*, page 3. In Russian.

Aiello, M. (2002a). *Spatial Reasoning: Theory and Practice*. PhD thesis, ILLC, University of Amsterdam. DS-2002-02.

Aiello, M. (2002b). A spatial similarity based on games: Theory and practice. *Journal of the Interest Group in Pure and Applied Logic*, 10(1):1–22.

Aiello, M. and Ottens, B. (2007). The mathematical morphological view on reasoning about space. In *International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 205–211. Morgan Kaufmann.

Aiello, M. and van Benthem, J. (2002a). Logical patterns in space. In Barker-Plummer, D., Beaver, D., van Benthem, J., and di Luzio, P. Scotto, editors, *Words, Proofs, and Diagrams*, pages 5–25. CSLI, Stanford.

Aiello, M. and van Benthem, J. (2002b). A modal walk through space. *Journal of Applied Non-Classical Logics*, 12(3–4):319–364.

Aiello, M., van Benthem, J., and Bezhanishvili, G. (2003). Reasoning about space: The modal way. *J. Logic Comput.*, 13(6):889–920.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843.

Allen, J. and Hayes, P. (1985). A common sense theory of time. In Joshi, A., editor, *IJCAI85*, volume 1, pages 528–531. International Joint Conference on Artificial Itelligence, Morgan Kaufmann.

Andréka, H., van Benthem, J., and Németi, I. (1998). Modal logics and bounded fragments of predicate logic. *Journal of Philosophical Logic*, 27(3):217–274.

Anger, F., van Benthem, J., Guesgen, H., and Rodriguez, R. (1996). Editorial of the special issue on 'Space, Time and Computation: Trends and Problems'. *International Journal of Applied Intelligence*, 6(1):5–9.

Areces, C. and ten Cate, B. (2006). Hybrid logics. In *Handbook of Modal Logic*.

Artemov, S. (2006). Modal logic in mathematics. In *Handbook of Modal Logic*.

Awodey, S. and Kishida, K. (2006). Topological semantics for first-order modal logic. In preparation.

Balbiani, Ph. (1998). The modal multilogic of geometry. *Journal of Applied Non-Classical Logics*, 8:259–281.

Balbiani, Ph., Fariñas del Cerro, L., Tinchev, T., and Vakarelov, D. (1997). Modal logics for incidence geometries. *Journal of Logic and Computation*, 7:59–78.

Barwise, J. (1988). Three views of common knowledge. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge (Pacific Grove, CA, 1988)*, pages 365–379, Los Altos, CA. Morgan Kaufmann.

Ben-Or, M., Kozen, D., and Reif, J. (1986). The complexity of elementary algebra and geometry. *Journal of Computer and System Sciences*, 32: 251–264.

Bennett, B. (1995). Modal logics for qualitative spatial reasoning. *Bulletin of the IGPL*, 3:1–22.

Bezhanishvili, G., Esakia, L., and Gabelaia, D. (2005). Some results on modal axiomatization and definability for topological spaces. *Studia Logica*, 81(3): 325–355.

Bezhanishvili, G. and Gehrke, M. (2005). Completeness of S4 with respect to the real line: revisited. *Ann. Pure Appl. Logic*, 131(1-3):287–301.

Bezhanishvili, G., Mines, R., and Morandi, P. (2003). Scattered, Hausdorff-reducible, and hereditary irresolvable spaces. *Topology and its Applications*, 132:291–306.

Bezhanishvili, N. and Kupke, C. (2006). Spatio-temporal logics of the real line. In preparation.

Binmore, K. (1994). *Game Theory and the Social Contract*. MIT Press, Cambridge.

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press.

Blass, A. (1990). Infinitary combinatorics and modal logic. *J. Symbolic Logic*, 55(2):761–778.

Bloch, I. (2000). Using mathematical morphology operators as modal operators for spatial reasoning. In *ECAI 2000, Workshop on Spatio-Temporal Reasoning*, pages 73–79.

Blumenthal, L. (1961). *A Modern View of Geometry*. Dover.

Boolos, G. (1993). *The Logic of Provability*. Cambridge University Press, Cambridge.

Burgess, J. (1984). Basic tense logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume II, chapter 2, pages 89–133. Reidel.

Burgess, John P. (1979). Logic and time. *J. Symbolic Logic*, 44(4):566–582.

Carnap, Rudolf (1998). *Der logische Aufbau der Welt*, volume 514 of *Philosophische Bibliothek [Philosophical Library]*. Felix Meiner Verlag, Hamburg. Reprint of the 1928 original and of the author's preface to the 1961 edition.

Chagrov, A. and Zakharyaschev, M. (1997). *Modal Logic*, volume 35 of *Oxford Logic Guides*. Clarendon Press, Oxford.

Chellas, B. (1980). *Modal Logic: An Introduction*. Cambridge University Press.

Dabrowski, A., Moss, A., and Parikh, R. (1996). Topological reasoning and the logic of knowledge. *Annals of Pure and Applied Logic*, 78:73–110.

de Jongh, D. and Veltman, F. (1985). Lecture Notes on Modal Logic. ILLC, Amsterdam.

de Rijke, M. (1993). *Extended Modal Logic*. PhD thesis, ILLC, University of Amsterdam.

Doets, Kees (1996). *Basic Model Theory*. Studies in Logic, Language and Information. CSLI Publications, Stanford, CA.

Engelking, R. (1989). *General Topology*. Heldermann Verlag.

Esakia, L. (1981). Diagonal constructions, Löb's formula and Cantor's scattered spaces. In *Studies in Logic and Semantics*, pages 128–143. Metsniereba. In Russian.

Esakia, L. (2001). Weak transitivity—restitution. In *Study in Logic*, volume 8, pages 244–254. Nauka. In Russian.

Esakia, L. (2002). The modal version of Gödel's second incompleteness theorem and the McKinsey system. In *Logical Investigations. Vol. IX*, pages 292–300. In Russian.

Esakia, L. (2004). Intuitionistic logic and modality via topology. *Annals of Pure and Applied Logic*, 127:155–170.

Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning About Knowledge*. MIT Press, Cambridge, MA.

Fontaine, G. (2006). Axiomatization of ML and Cheq. Master's thesis, ILLC, University of Amsterdam.

Gabbay, D., Kurucz, A., Wolter, F., and Zakharyaschev, M. (2003). *Many-Dimensional Modal Logics: Theory and Applications*. Elsevier, Uppsala. Studies in Logic and the Foundations of Mathematics, Volume 148.

Gabbay, D. and Shehtman, V. (1998). Products of modal logics. I. *Log. J. IGPL*, 6(1):73–146.

Gabelaia, D. (1999). Modal logics GL and Grz: semantical comparison. In *Proceedings of the ESSLLI Student Session*, pages 91–97.

Gabelaia, D. (2001). Modal Definability in Topology. Master's thesis, ILLC, University of Amsterdam.

Gabelaia, D. (2004). *Topological, Algebraic and Spatio-Temporal Semantics for Multi-Dimensional Modal Logics*. PhD thesis, King's College, London.

Gabelaia, D., Kurucz, A., Wolter, F., and Zakharyashev, M. (2005). Products of 'transitive' modal logics. *Journal of Symbolic Logic*, 70:993–1021.

Gabelaia, D. and Sustretov, D. (2005). Modal correspondence for topological semantics. In *Abstracts of Algebraic and Topological Methods in Non-Classical Logics II (Barcelona 2005)*, pages 80–81.

Gärdenfors, P. (2000). *Conceptual Spaces*. MIT Press.

Gerhardt, S. (2004). A Construction Method for Modal Logics of Space. Master's thesis, ILLC, University of Amsterdam.

Goldblatt, R. (1980). Diodorean modality in Minkowski space-time. *Studia Logica*, 39:219–236.

Goldblatt, R. (1987). *Orthogonality and Spacetime Geometry*. Springer-Verlag.

Goranko, V. and Passy, S. (1992). Using the universal modality: Gains and questions. *J. Logic Comput.*, 2(1):5–30.

Harel, D., Kozen, D., and Tiuryn, J. (2000). *Dynamic Logic*. Foundations of Computing Series. MIT Press, Cambridge, MA.

Helly, E. (1923). Über Mengen konvexer Körper mit gemeinschaftlichen Punkten. *Jahresber. Deutsch. Math. Verein*, 32:175–176.

Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press.

Hodkinson, I. and Reynolds, M. (2006). Temporal logic. In *Handbook of Modal Logic*.

Kamp, J. (1968). *Tense Logic and the Theory of Linear Order*. PhD thesis, University of California, Los Angeles.

Kelley, J. (1975). *General Topology*. Springer-Verlag, New York.

Konev, B., Kontchakov, R., Wolter, F., and Zakharyaschev, M. (2004). On dynamic topological and metric logics. In *Proceedings of AiML 2004*, pages 182–196. King's College Publications.

Kuratowski, K. (1966). *Topology. Vol. I*. Academic Press, New York.

Kuratowski, K. and Mostowski, A. (1976). *Set Theory*. North Holland, Amsterdam-New York-Oxford.

Kurtonina, N. (1995). *Frames and Labels. A Modal Analysis of Categorial Inference*. PhD thesis, ILLC, Amsterdam.

Kurtonina, N. and de Rijke, M. (1997). Bisimulations for temporal logic. *Journal of Logic, Language and Information*, 6:403–425.

Leitgeb, H. (2007). 'A new analysis of quasianalysis'. *Journal of Philosophical Logic*.

Lewis, D. (1969). *Convention*. Harvard University Press.

Litak, T. (2004). Some notes on the superintuitionistic logic of chequered subsets of R^∞. *Bull. Sect. Logic Univ. Łódź*, 33(2):81–86.

Marx, M. (1995). *Algebraic Relativization and Arrow Logic*. PhD thesis, ILLC.

Marx, M. and Venema, Y. (1997). *Multi Dimensional Modal Logic*. Kluwer.

Matheron, G. (1967). *Eléments pur Une Theorie des Milieux Poreaux*. Masson.

McKinsey, J. and Tarski, A. (1944). The algebra of topology. *Annals of Mathematics*, 45:141–191.

Mikulas, S. (1995). *Taming Logics*. PhD thesis, ILLC.

Mints, G. (1998). A completeness proof for propositional S4 in Cantor Space. In E. Orlowska, editor, *Logic at work : Essays dedicated to the memory of Helena Rasiowa*. Physica-Verlag, Heidelberg.

Moerdijk, I. (1982). Some topological spaces which are universal for intuitionistic predicate logic. *Indagationes Mathematicae*, 44(2):227–235.

Pauly, M. (2001). *Logic for Social Software*. PhD thesis, ILLC, University of Amsterdam.

Peleg, D. (1987). Concurrent dynamic logic. *Journal of the ACM*, 34(2): 450–479.

Pratt, V. (1999). Chu spaces. In *School on Category Theory and Applications (Coimbra, 1999)*, volume 21 of *Textos Mat. Sér. B*, pages 39–100. Univ. Coimbra, Coimbra.

Preparata, F. and Shamos, M. (1985). *Computational Geometry: An Introduction*. Springer-Verlag.

Randell, D., Cui, Z., and Cohn, A. (1992). A spatial logic based on regions and connection. In *Proc. of Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 165–176. San Mateo.

Randell, D., Witkowski, M., and Shanahan, M. (2001). From images to bodies: Modelling and exploiting occlusion and motion parallax. In *Proc. of Int. Joint Conference on Artificial Intelligence (IJCAI-01)*.

Rasiowa, H. and Sikorski, R. (1963). *The Mathematics of Metamatematics*. Panstwowe Wydawnictwo Naukowe.

Segerberg, K. (1970). Modal logics with linear alternative relations. *Theoria*, 36:301–322.

Segerberg, K. (1973). Two-dimensional modal logic. *J. Philos. Logic*, 2(1): 77–96.

Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press.

Shehtman, V. (1983). Modal logics of domains on the real plane. *Studia Logica*, 42:63–80.

Shehtman, V. (1990). Derived sets in Euclidean spaces and modal logic. Technical Report X-1990-05, Univ. of Amsterdam.

Shehtman, V. (1993). A logic with progressive tenses. In *Diamonds and defaults (Amsterdam, 1990/1991)*, volume 229 of *Synthese Lib.*, pages 255–285. Kluwer Acad. Publ., Dordrecht.

Shehtman, V. (1999). 'Everywhere' and 'here'. *Journal of Applied Non-Classical Logics*, 9(2-3):369–379.

Shehtman, V. (2006). Derivational modal logics. *Moscow Mathematical Journal*. submitted.

Sheremet, M., Tishkovsky, D., Wolter, F., and Zakharyaschev, M. (2006). Comparative similarity, tree automata, and diophantine equations. *LPAR*. to appear.

Spaan, E. (1993). *Complexity of Modal Logics*. PhD thesis, University of Amsterdam, Institute for Logic, Language and Computation.

Stebletsova, V. (2000). *Algebras, Relations and Geometries*. PhD thesis, University of Utrecht.

Stebletsova, V. and Venema, Y. (2001). Undecidable theories of Lyndon algebras. *J. Symbolic Logic*, 66(1):207–224.

Steinsvold, C. (2005). Personal communication.

Szczerba, L. and Tarski, A. (1965). Metamathematical properties of some affine geometries. In Bar-Hillel, Y., editor, *Int. Congress for Logic, Methodolog, and Philosophy of Science*, pages 166–178. North-Holland.

Tarski, A. (1938). Der Aussagenkalkül und die Topologie. *Fund. Math.*, 31: 103–134.

Tarski, A. (1959). What is elementary geometry? In L. Henkin and P. Suppes and A. Tarski, editor, *The Axiomatic Method, with Special Reference to Geometry ad Physics*, pages 16–29. North-Holland.

ten Cate, B., Gabelaia, D., and Sustretov, D. (2006). Modal languages for topology: expressivity and definability. Preprint.

Troelstra, A. (1992). *Lectures on Linear Logic*. CSLI.

van Benthem, J. (1983a). Correspondence theory. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume II, pages 167–247. Reidel Publishing Company.

van Benthem, J. (1983b). *The Logic of Time*, volume 156 of *Synthese Library*. Reidel, Dordrecht. [Revised and expanded, Kluwer, 1991].

van Benthem, J. (1991a). *Language in Action. Categories, Lambdas and Dynamic Logic*. North-Holland, Amsterdam.

van Benthem, J. (1991b). Logic and the flow of information. In Prawitz, D., Skyrms, B., and Westertal, D., editors, *Proceedings of the 9th International Conference of Logic, Methodology and Philosophy of Science*, pages 693–724. Elsevier.

van Benthem, J. (1992). Logic as programming. *Fundamenta Informaticae*, 17(4):285–317.

van Benthem, J. (1995). Temporal logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming, Vol. 4*, Oxford Sci. Publ., pages 241–350. Oxford Univ. Press, New York.

van Benthem, J. (1996). *Exploring Logical Dynamics*, volume 156. CSLI Publications, Stanford/Cambridge University Press.

van Benthem, J. (1999). Temporal patterns and modal structure. *Log. J. IGPL*, 7(1):7–26. Special issue on Temporal Logic. A. Montanari, A. Policriti, and Y. Venema eds.

van Benthem, J. (2000a). Information transfer across Chu spaces. *Log. J. IGPL*, 8(6):719–731.

van Benthem, J. (2000b). Logical structures in mathematical morphology. Available at http://www.science.uva.nl/ johan/MM-LL.ps.

van Benthem, J. (2002). Invariance and definability: two faces of logical constants. In Sieg, W., Sommer, R., and Talcott, C., editors, *Reflections on the Foundations of Mathematics. Essays in Honor of Sol Feferman*, ASL Lecture Notes in Logic, pages 426–446. ASL.

van Benthem, J., Bezhanishvili, G., Cate, B. ten, and Sarenac, D. (2005). Modal logics for products of topologies. *Studia Logica*, 84(3):369–392.

van Benthem, J., Bezhanishvili, G., and Gehrke, M. (2003). Euclidean hierarchy in modal logic. *Studia Logica*, 75(3):327–344.

van Benthem, J. and Blackburn, P. (2006). Basic modal model theory. In *Handbook of Modal Logic*.

van Benthem, J. and Sarenac, D. (2004). The geometry of knowledge. In *Aspects of universal logic*, volume 17 of *Travaux Log.*, pages 1–31. Univ. Neuchâtel, Neuchâtel.

van Dalen, Dirk (2005). *Mystic, Geometer, and Intuitionist*. The Clarendon Press Oxford University Press, Oxford. The life of L. E. J. Brouwer, 1881–1966. Vol. 2, Hope and disillusion.

Venema, Y. (1996). A crash course in arrow logic. In Marx, M., Masuch, M., and Pólos, L., editors, *Arrow Logic and Multimodal Logic*. CSLI.

Venema, Y. (1999). Points, lines and diamonds: A two-sorted modal logic for projective planes. *Journal of Logic and Computation*, 9(5):601–621.

Venema, Y. (2007). Algebras and Co-algebras. In Blackburn, P., van Benthem, J. and Wolter, F., editors, Handbook of Model Logic, pages 331–426. Elsevier, Amsterdam.

von Plato, J. (1995). The axioms of constructive geometry. *Annals of Pure and Applied Logic*, 76(2):169–200.

Wooldridge, M. (2002). *An Introduction to Multiagent Systems*. J. Wiley, New York.

Chapter 6

TOPOLOGY AND EPISTEMIC LOGIC

Rohit Parikh
City University of New York

Lawrence S. Moss
Indiana University

Chris Steinsvold
City University of New York

Second Reader

Darko Sarenac
Stanford University

1. Introduction

This handbook chapter explores some themes which relate general topology and epistemic logic. The leading ideas are: (1) to review the connection between the modal logic $S4$ and topology going back to the work of Alfred Tarski and J. C. C. McKinsey in the 1940's; (2) to discuss the epistemic interpretation of topology; (3) to present the two-sorted semantics of topologic and to discuss what is known about it, including some of the main completeness and decidability results; (4) to present a topological semantics for the logic of belief $KD45$ based on the derived set operation; and (5) to briefly mention related work in a number of directions.

Topology and modal logic: a first look. One of the things which strikes one when studying elementary (set-theoretic) topology is how easy it is. Notions like *open, closed, dense*, seem intuitively transparent: their basic properties easy to prove. Contrasting this fact is that topology uses second order notions

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 299–341.
© 2007 *Springer.*

as it reasons with both points and sets. This would imply that like second order logic, topology ought to be computationally very difficult.

This intuitive tension between the two paradigms vanishes when one realizes that a large part of topology can be seen as a *modal logic*, i.e., as an epistemic logic which combines the notion of knowledge and effort. Recall that modal logics tend to be much easier than first order logic, let alone second order.

Suppose that we have made a *measurement*—for example, that some velocity v is $50 \pm .5$. We interpret this as saying that v is in the open set (49.5,50.5) and therefore anything which we *know* about v must hold not only of v itself, but also of any v' in the same interval. (49.5,50.5) thus becomes an equivalence class for an appropriate $S5$ logic of knowledge. Thus the connection with modal logic.

However, the notion of *effort* can also enter, as v might be measured more accurately with more effort. Following up on this basic intuition, the first two authors of this chapter developed a bimodal logic called topologic for studying elementary topology. This logic turned out to have a nice axiomatization and to be decidable. The original work was followed up then by further work by Georgatos and also by Dabrowski in conjunction with the two original writers (see Georgatos, 1993; Georgators, 1994a; Georgatos, 1994b; Georgatos, 1997; also, Heinemann, 1997; Heinemann, 1999b; Heinemann, 2001; Weiss, 1999) We should mention that the original logic was defined for arbitrary subset spaces. The subsequent logics consider extensions of this logic where restricted families of sets were considered, closed under union or intersection. The bulk of our chapter, Sections. 4–8, is about topologic.

Another large part of the chapter comes from Steinsvold's discovery that the notion of *belief* can also be given a topological meaning via the notion of derived set. This is covered in Sec. 10; the main reference is his dissertation (Steinsvold, 2006).

This chapter may be considered a continuation of van Benthem and Bezhanishvili's Ch. 5 of this handbook. Although we have written this chapter to be read on its own, readers would certainly benefit from a look at the many related discussions about different logics that can be found in Ch. 5.

2. Perspectives

This handbook deals with *logic* and *space*, more precisely with various logics formulated with different goals in mind. But in a certain sense, the goal is usually to take a common mathematical model of space and then to fashion logical tools to work with it. This chapter works in a different way. The overall points are to investigate notions such as *knowledge*, *belief*, and *observation effort*; and mathematical structures like topological spaces arise in the course of that investigation. So the modus operandi, and the overall goals, of our work

are different. One should not expect insight into, say, algebraic topology from our work. What we are after is rather a kind of reconstruction of the ideas underlying topology.

Some perspective on our subject might be gleaned from comparing the ideas with those in a well-known source, Steven Vickers' book *Topology via Logic* (see Vickers, 1989). One goal of that book is to explain the non-Hausdorff topologies that arise in computer science. This goal is relevant for our chapter, more so perhaps than for other chapters of this handbook. The overall message is that (p. 3) "topology is used to explain *approximate* states of information: the points include both approximate points and more refined points, and these relate to the topology by the property that if an open set contains an approximate point, then it must contain any refinement of it." The book goes on (pp. 5–11) to discuss "finite observations," and therefore aims at a reconstruction of topology in terms of logics of observations. The algebraization of a logic of finite observations is called a *frame*; it is a complete Heyting algebra under two operations $\wedge$ (binary) and $\bigvee$ (infinitary). (Note that the opens of a topological space are a frame under the corresponding set-theoretic operations.) This notion of a frame gives us a more topological notion, a *locale*; this is based on the frame morphisms into a special two-element frame.

Now our work differs from Vickers' in the sense that we add to the notion of observation a notion of *effort*. Our discussion begins in Sec. 4. Our treatment of effort is in a sense fairly crude: we take the subsets of a space as the possible observations, and then more effort corresponds to a smaller set, a better approximation to being closer to the "real" point. We say that this is crude because it does not measure the amount of effort in any real sense. However, it is a refinement of treatments which do not include any modeling of effort whatsoever. We start in the next section by first discussing the classical work of Tarski and McKinsey.

3. The original topological interpretation of modal logic: Tarski and McKinsey's Theorem

The project of relating topology to modal logic begins with work of Alfred Tarski and J.C.C. McKinsey (see McKinsey, 1941; McKinsey, 1944). The basic idea is to study the laws of the *interior* operation on subsets of a topological space and its dual, the *closure* operator. Suppose that $\mathcal{X} = \langle X, \mathcal{O} \rangle$ is a *topological space*; this just means that $\mathcal{O}$ is a family of subsets of X containing the empty set $\emptyset$ and X itself, and $\mathcal{O}$ is closed under arbitrary unions and finite intersections. The family $\mathcal{O}$ is called a *topology*, and its elements are the *open sets* of the space $\mathcal{X}$. The *closed* sets are the complements of the opens. For any subset $A \subseteq X$, we define its interior A^{O} to be the largest open subset of A. Dually, the closure $\overline{A}$ is the smallest closed set including A. We say "dually"

here to illustrate one of the properties of these operations: $A^{\mathrm{o}} = X \setminus (\overline{X \setminus A})$. In words, the interior of A is the complement of the closure of the complement of A. This holds for all subsets of all spaces; it is just this kind of general property that we aim to study.

In order to study these notions, we introduce a logical language $\mathcal{L}_0$. It is a modal language. We begin with an arbitrary but fixed set At of atomic propositions and close under truth functions $\wedge$ and $\neg$ and the appropriate modalities; for present purposes, we use the modality I for the interior operation. (As we have seen, I and C are interdefinable duals. So only one needs to be taken as basic in the syntax, and then the other may be regarded as a defined symbol.)

The language $\mathcal{L}_0$ is interpreted on such a space $\mathcal{X}$ together with an *interpretation map into* $\mathcal{X}$, $i : At \to \mathcal{P}(X)$. (We do not insist that each $i(p)$ is an open set.) So for atomic p, $i(p)$ says which points satisfy p. We call $\langle X, \mathcal{O}, i\rangle$ a *topological model*. Then i extends to all of $\mathcal{L}_0$ by interpreting negation as complement relative to X, conjunction as intersection, and C and I as the closure and interior operators respectively. In symbols, we have

$$\begin{array}{lcl} i(\neg\phi) & = & X \setminus i(\phi) \\ i(\phi \wedge \psi) & = & i(\phi) \cap i(\psi) \\ i(I\phi) & = & (i(\phi))^{\mathrm{o}} \end{array}$$

EXAMPLE 6.1 For a very easy example, consider the usual real line $\mathcal{R}$ with $i(p) = \{1\}$. Then $i(Ip) = \emptyset$, $i(\neg p) = R\setminus\{1\}$, and $i(I\neg p) = i(\neg p)$. Moreover, one can check as well that for all ϕ, $i(\phi)$ is always one of the sets $\emptyset$, R, $\{1\}$, or $R \setminus \{1\}$.

As with all attempts to study some phenomenon, the main idea here is that the basic properties of the boolean operations on sets (unions, intersection, and complement), and also the salient topological operations (interior and closure) correspond to sentences in the language, or rather to schemes of sentences. For example, consider another general fact: the interior of the intersection of two sets is the intersection of their interiors. This corresponds to the fact that for all sentences ϕ and ψ, and for all spaces $\mathcal{X}$ and all interpretations i of whatever atomic sentences we have, we also have

$$i((I(\phi \wedge \psi) \leftrightarrow ((I\phi) \wedge (I\psi))) \quad = \quad X.$$

For another example, the fact that the interior operation is idempotent corresponds in the same way to the scheme

$$i((II\phi) \leftrightarrow (I\phi)) \quad = \quad X.$$

One of the natural questions to ask about this language and its semantics is: can we characterize in an enlightening way the sentences ϕ with the property

that for all topological models $\langle X, \mathcal{O}, i\rangle$, $i(\phi) = X$? We call such sentences *topologically valid*. The reason for this is that we more generally write $x \models \phi$ for $x \in i(\phi)$. (This notation from model theory will be used throughout our chapter.) Then the topologically valid sentences are those that are true at all points in all spaces under all interpretations. Similarly, the *topologically satisfiable* sentences are those sentences ϕ which belong to $i(\phi)$ for some topological model. And we say that a set $S \subseteq \mathcal{L}$ is topologically satisfiable if there is a topological model with $\bigcap_{\phi \in S} i(\phi) \neq \emptyset$.

What Tarski and McKinsey proved is that the topologically valid sentences are exactly those provable in the logical system $S4$. This is a logical system which had been proposed much earlier than their work. It is the system whose axioms are the substitution instances of the tautologies of classical propositional logic, and also the schemes below:

1 $I(\phi \to \psi) \to (I\phi \to I\psi)$

2 $I\phi \to \phi$

3 $I\phi \to II\phi$

(Note that the second of these can be read as saying that the interior of a set is a subset of it, and the last is one direction of the idempotence of the interior operation.) The rules of $S4$ are modus ponens, and also necessitation: from ϕ, derive $I\phi$. It should be mentioned that the earliest work on $S4$ did not present this semantics, so that the Tarski-McKinsey work may be read as giving a nice semantics to an already-existing logical system.

THEOREM 6.2 (TARSKI AND MCKINSEY) *The interpretation of $\mathcal{L}_0$ in topological spaces is sound and complete in the sense that the following are equivalent:*

1 ϕ is topologically valid.

2 ϕ is provable in $S4$.

Moreover, the set of topologically valid sentences is decidable.

Proof The soundness being routine, here is a sketch of the completeness. Our work here follows a recent presentation (Aiello et al., 2003). One considers the *theories* in $S4$; these are the maximal consistent sets of sentences. We show that each theory T is topologically satisfiable. This is equivalent to completeness. Actually, we shall show that all theories T are satisfiable in the same *canonical topological model*.

Let $\mathcal{C}(S4)$ be the set of theories in $S4$. For any ϕ, let

$$\widehat{\phi} \quad = \quad \{T \in \mathcal{C} : \phi \in T\}.$$

The family of sets $\{\widehat{I\phi} : \phi \in \mathcal{L}_0\}$ is the basis of a topology; this is tantamount to the fact about intersections and interiors that we noted above. In this topology, the basic open sets are those of the form $\{T : I\phi \in T\}$, for some sentence ϕ. We use the following *canonical interpretation*

$$i(p) \quad = \quad \widehat{p} \tag{6.1}$$

for $p \in At$. This gives a topological model which we also call $\mathcal{C}(S4)$.

The main fact about $\mathcal{C}(S4)$ is the following Truth Lemma:

$$i(\phi) \quad = \quad \widehat{\phi}.$$

The proof is by induction on ϕ. The base case for atomic sentences is by definition, and the steps for the propositional connectives use basic facts about theories. The main work is in the induction step for $I\phi$. Suppose first that $T \in \widehat{I\phi}$, so that $I\phi \in T$. Then this set $\widehat{I\phi}$ is a basic open set containing T. By Scheme 2 of the logic, $\widehat{I\phi} \subseteq \widehat{\phi}$. By induction hypothesis, $\widehat{\phi} = i(\phi)$. In this way, $T \in i(\phi)^{\mathrm{o}} = i(I\phi)$.

Going the other way, suppose that $T \in i(I\phi)$. Then there is some basic open set around T, say $\widehat{I\psi}$, included in $i(\phi)$. By induction hypothesis, $\widehat{I\psi} \subseteq \widehat{\phi}$. We claim now that $I\psi \to \phi$ is provable in $S4$. (If not, then $I\psi \wedge \neg\phi$ is consistent. So there is some theory $U \in i(I\psi)$ but not in $i(\phi)$. This contradicts $\widehat{I\psi} \subseteq \widehat{\phi}$.) Using this claim and the necessitation rule, we can prove $I(I\psi \to \phi)$. Using Scheme 1, we get $II\psi \to I\phi$. By Scheme 3, we then get $I\psi \to I\phi$. So $I\phi$ belongs to our original T. Hence $T \in \widehat{I\phi}$, as desired.

The final assertion of our theorem, the decidability, is a standard *effective finite model* result. The idea is that given a sentence ϕ, one can compute a number $n = n(\phi)$ such that if ϕ is satisfiable on any model, then it is satisfied on a model of size at most n. We get such a number by estimating the size of a certain quotient of $\mathcal{C}(S4)$, a quotient which of course depends on ϕ. We omit the details here, but see Sec. 8 for another decidability result. QED

There are also stronger forms of Theorem 6.2. For example, one might well wonder whether $S4$ is complete for the topological operations on specific natural spaces, such as the reals or the interval $[0, 1]$. McKinsey and Tarski showed that $S4$ is complete for every dense-in-itself separable metric space (see McKinsey, 1944). This implies the completeness of $S4$ for all of the spaces mentioned above. In recent years, there has been a series of papers simplifying the completeness arguments for these special cases; see, for example, Aiello et al., 2003.

3.1 The preorder semantics of the same system $S4$

We have already seen the topological interpretation of the logic $S4$. In what follows, we need not only this semantics but the more standard *relational*, or

Kripke semantics of this and other logical systems. This section reviews this topic.

The syntax of modal logic is similar to what we have already seen. It begins with some set At of *atomic sentences (or propositions)* and then considers the closure of this set under the boolean operations of $\wedge$ and $\neg$ (and others, say by abbreviation), and some modal operators. We shall continue to use I as the one operator for now. (It is more standard to use symbols like $\Box$ and K for the operators, with duals $\Diamond$ and L. We shall see these in later parts of our chapter.)

The semantics begins with a *frame*, a set whose elements are called *worlds* or *points* together with a binary relation on it. This relation is sometimes called *accessibility*, and symbols like $\rightarrow$, $\leq$, or R have been used for it. To interpret $S4$, read I not as interior but rather as "all points which the current point relates to". Then reading the schemes of $S4$ this way suggests that the accessibility relation be a *preorder* (transitive and reflexive): for example, $I\phi \rightarrow \phi$ says that if something is true of all points which the current point relates to, then it is true at the current point itself. To get a sound interpretation of $S4$, we should require that the current point is related to itself. We therefore define a *preorder model* to be a triple $\mathcal{X} = \langle X, \leq, i\rangle$, where $\langle X, \leq\rangle$ is a preorder and again $i : At \rightarrow \mathcal{P}(X)$.

The semantics is the same as for topological models, except that now the clause for I becomes

$$i(I\phi) \qquad = \qquad \{x : \uparrow x \subseteq i(\phi)\}$$

Here $\uparrow x$ stands for $\{y : x \leq y\}$. The semantics of C is given by duality, so that $C\phi \leftrightarrow \neg I \neg \phi$ is valid by definition. We again define *validity* and *satisfiability in preorder models* just as with topological models, mutatis mutandis.

THEOREM 6.3 *The interpretation of $\mathcal{L}_0$ in preorders is sound and complete in the sense that the following are equivalent:*

1 *ϕ is valid in preorder models.*

2 *ϕ is provable in $S4$.*

One can prove this result in the same manner as Theorem 6.2. We use the set $\mathcal{C}(S4)$ of theories in $S4$ and define a preorder on them by

$$T \leq U \text{ iff } \phi \in U \text{ whenever } I\phi \text{ in } T. \tag{6.2}$$

One also uses the canonical interpretation from (6.1). The rest of the completeness argument is standard, and any textbook on modal logic would contain it. Instead of giving the details, we present an alternative approach based on the connection between certain topological spaces and preorders.

Let $\mathcal{X} = \langle X, \leq \rangle$ be a preorder. Consider the *Alexandrov topology* on $\mathcal{X}$: the opens are the sets closed upwards in the order. This gives a topology which we call $\mathcal{O}_{\leq}$. This associates topological spaces to preorders. (Actually, the reflexivity is not used in verifying that we have a topology, a fact which will come into play in Sec. 11.2.) It also associates topological models to preorder models, just by copying the interpretation i. (Actually, this gives a very special kind of space: the opens are closed under arbitrary intersections.)

PROPOSITION 6.4 *For all preorder models $\langle X, \leq, i \rangle$, all $x \in X$, and all $\phi \in \mathcal{L}_0$,*

$$x \models \phi \textit{ in } \langle X, \leq, i \rangle \quad \textit{iff} \quad x \models \phi \textit{ in } \langle X, \mathcal{O}_{\leq}, i \rangle$$

Proof By induction on $\phi \in \mathcal{L}_0$. The case of the atomic sentences is trivial, as are the induction steps for the boolean connectives.

Assume the lemma for ϕ. So $i(\phi)$ is the same for both interpretations. Consider $I\phi$. First, assume that $x \in i(I\phi)$ in the preorder sense. That is, every $y \geq x$ is in $i(\phi)$. Now $\uparrow x$ is an open set, and by reflexivity it contains x. As we have just seen, it is included in $i(\phi)$. So $x \in i(I\phi)$ in the topological sense. Conversely, if $x \in I(\phi)$ topologically, then there is some y such that $x \in \uparrow y$ and $\uparrow y \subseteq i(\phi)$. But then $y \leq x$. By transitivity, $\uparrow x \subseteq \uparrow y$. So $\uparrow x \subseteq i(\phi)$. Thus $x \in i(I\phi)$ in the preorder sense. QED

Using this easy result, Theorem 6.2 follows from Theorem 6.3. That is, every $S4$ theory is satisfied on some preorder, hence on some topological space. But with a little more work, Theorem 6.3 can be made to follow from Theorem 6.2 or rather from the related fact that sentences which are satisfiable in $S4$ have *finite* topological models. Thus they are *Alexandrov spaces*: every point has a minimal open set around it. We turn a topological model which is an Alexandrov space into a preorder model by: $x \leq y$ iff y belongs to every open set around x: the Alexandrov property implies the transitivity. Then one notes a result just like Proposition 6.4, except now we relate the topological semantics on an Alexandrov space to the preorder semantics derived from it. The upshot is that a given topologically satisfiable sentence ϕ now has a finite preorder model. This implies the completeness of $S4$ in the relational semantics.

3.2 Adding the difference modality

One of the purposes of this chapter is to mention other work that builds on classical topics. We mention here recent work (Gabelaia, 2001; Kudinov, 2006) which adds a modal operator to the language we have been discussing. Kudinov's paper is the source for all results in this subsection and contains other material as well. Add to $\mathcal{L}_0$ the operator $[\neq]$ to get a larger language $\mathcal{L}_1$. For

the semantics, we stipulate that

$$x \models [\neq]\phi \quad \text{iff} \quad \text{for all } y \text{ different from } x\text{, we have } y \models \phi.$$

So we can read $[\neq]\phi$ as "ϕ holds everywhere except possibly here." It is also useful to adopt an abbreviation $K\phi$ for $\phi \wedge [\neq]\phi$, (Kudinov uses $[\forall]$, but we use K later for just this purpose). As usual we include for convenience a dual modality $\langle\neq\rangle$ so that $\langle\neq\rangle\phi$ abbreviates $\neg[\neq]\neg\phi$. Then as axioms, one takes $S4$ for I (just as we have seen), together with the following schemes:

$$\begin{array}{l}
[\neq](\phi \to \psi) \to ([\neq]\phi \to [\neq]\psi) \\
\phi \to [\neq]\langle\neq\rangle\phi \\
(\phi \wedge [\neq]\phi) \to [\neq][\neq]\phi \\
K\phi \to I\phi
\end{array}$$

For rules we take the necessitation rules for both I and also $[\neq]$. This axiom system is called $S4D$. It is easy to see that $S4D$ is sound for all topological interpretations.

Now in this logic we can express some interesting topological properties. The examples we have in mind are: density-in-itself, and T_1. The first condition means that there are no isolated points: $\{x\}$ is not open. The second means that for every $x \neq y$, there is an open set containing x but not y.) These are examples of *correspondence phenomena*, and the formal statements involve quantification over interpretations of the atomic sentences in the model, in the following way.

PROPOSITION 6.5 *Let $\mathcal{X} = \langle X, \mathcal{O}\rangle$ be a topological space.*

1 *The following are equivalent:*

   (a) *$\mathcal{X}$ is dense in itself.*

   (b) *For all sentences ϕ, all interpretations i into $\mathcal{X}$, and all $x \in X$, we have $x \models [\neq]\phi \to C\phi$.*

2 *The following are equivalent:*

   (a) *$\mathcal{X}$ is a T_1 space.*

   (b) *For all sentences ϕ, all interpretations i into $\mathcal{X}$, and all $x \in X$, we have $x \models [\neq]\phi \to [\neq]I\phi$.*

Correspondence results for the T_0 and T_1 properties were first obtained in Gabelaia, 2001. The main results about the logics in Proposition 6.5 are the following completeness theorems. For example, the second statement below means that for a sentence $\phi \in \mathcal{L}_1$, ϕ is provable in the system axiomatized

by $S4D$ with the extra scheme $[\neq]\phi \to C\phi$ iff ϕ holds in every space $\mathcal{X}$ which is dense in itself, at all points, under all valuations.

THEOREM 6.6 (KUDINOV, 2006)

1 *$S4D$ is complete for topological spaces.*

2 *$S4D$ + "$[\neq]\phi \to C\phi$" is complete for spaces which are dense in themselves.*

3 *$S4D$ + "$[\neq]\phi \to C\phi$" + "$[\neq]\phi \to [\neq]I\phi$" is complete for spaces which are dense in themselves and also T_1.*

4. Topologic

At this point we have seen the topological interpretation of modal logic. This topic has been pursued in many directions over the years; see Ch. 5 for a survey. It is not exactly the thrust of this chapter, however. Instead, we strike out on a different direction by considering a *bimodal* language interpreted on a larger class of models. This language was first considered in Moss and Parikh, 1992.

A *subset frame* is a pair $\mathcal{X} = \langle X, \mathcal{O}\rangle$ where X is a set of *points* and $\mathcal{O}$ is a set of subsets of X. The elements of $\mathcal{O}$ are called *opens*. We assume that $X \in \mathcal{O}$, though this is really not necessary. $\mathcal{X}$ is an *intersection frame* if whenever $u, v \in \mathcal{O}$ and $u \cap v \neq \emptyset$, then also $u \cap v \in \mathcal{O}$. $\mathcal{X}$ is a *lattice frame* if it is an intersection frame closed under finite unions, and a *complete lattice frame* if it is closed under infinitary intersections and unions.

Note that we use the term "open" for simplicity even though it is not required that $\mathcal{O}$ be a topology. It is just that the topological case is our paradigm case, and the basis of our intuitions.

We now set up a formal language $\mathcal{L}$ which is expressive enough for simple arguments concerning subset spaces. Later we shall expand this language. The formulas of $\mathcal{L}$ are obtained from atomic propositions by closing under $\wedge$, $\neg$, K and $\Box$.

A *subset space* is a triple $\mathcal{X} = \langle X, \mathcal{O}, i\rangle$, where $\langle x, \mathcal{O}\rangle$ is a subset frame, and $i : At \to \mathcal{P}(X)$. (We do not require that each $i(p)$ be an open.) If $\langle x, \mathcal{O}\rangle$ is an intersection frame, then $\mathcal{X}$ is called an *intersection space*, and similarly for lattice spaces, etc. (Often we simply speak of *models*.) For $p \in X$ and $p \in u \in \mathcal{O}$, we define the *satisfaction relation* $\models_{\mathcal{X}}$ on $(X \times \mathcal{O}) \times \mathcal{L}$ by recursion on ϕ.

$p, u \models_{\mathcal{X}} A$	iff	$p \in i(A)$
$p, u \models_{\mathcal{X}} \phi \wedge \psi$	iff	$p, u \models_{\mathcal{X}} \phi$ and $p, u \models_{\mathcal{X}} \psi$
$p, u \models_{\mathcal{X}} \neg\phi$	iff	$p, u \not\models_{\mathcal{X}} \phi$
$p, u \models_{\mathcal{X}} K\phi$	iff	$q, u \models_{\mathcal{X}} \phi$ for all $q \in u$
$p, u \models_{\mathcal{X}} \Box\phi$	iff	$p, v \models_{\mathcal{X}} \phi$ for all $v \in \mathcal{O}$ such that $p \in v \subseteq u$.

We use L as the dual of K and $\Diamond$ as the dual of $\Box$. Explicitly, we have

$$\begin{array}{lll} p, u \models_{\mathcal{X}} L\phi & \text{iff} & q, u \models_{\mathcal{X}} \phi \text{ for some } q \in u \\ p, u \models_{\mathcal{X}} \Diamond\phi & \text{iff} & p, v \models_{\mathcal{X}} \phi \text{ for some } v \in \mathcal{O} \text{ such that } p \in v \subseteq u. \end{array}$$

We stress that we only use the notation $p, v \models \phi$ when p belongs to v.

As usual, we write $p, u \models \phi$ if $\mathcal{X}$ is clear from context. If $T \subseteq \mathcal{L}$, we write $T \models \phi$ if for all models $\mathcal{X}$, all $p \in X$, and all $u \in \mathcal{O}$, if $p, u \models_{\mathcal{X}} \psi$ for each $\psi \in T$, then also $p, u \models_{\mathcal{X}} \phi$. Finally, we also write, e.g., $T \models_{Int} \phi$ for the natural restriction of this notion to the class of models which are intersection spaces.

With these definitions in place, we return to a discussion of the concepts and motivation. We are considering a Kripke structure whose worlds are the pairs (p, u) with $p \in u$ and $u \in \mathcal{O}$. Think of p as the "real world" and u as a guess as to where that world lies based on some observation. The language then uses two accessibility relations corresponding to *shrinking* an open ($\Box$) while maintaining a reference point, or to moving a reference point inside the *given open* (K). Another way to think of these is in terms of quantification, so we recast the semantics in English a bit: $\Box\phi$ is true at p, u if for all refinements v of the observation u, ϕ is still true at p, v. And $K\phi$ is true at p, u if for all $q \in u$, if the real world happens to be q rather than p, ϕ is still true at that world q with the same observation u.

We see that the intuition behind this logic with its two modalities is that knowledge is affected not only by the *situation* we are in, but also by the amount of *effort* we have put in. For instance, recall our remarks introducing measurement in the Introduction. Suppose a policeman uses radar to determine that a car is going 51 mph in a 50 mile speed limit zone. But if the accuracy of his radar is ± 2 mph, then he does not know that the car is speeding. If however, a more accurate radar with an accuracy of ± 1 mph shows that the car is going 51.5 mph, then he *does know* that the car is speeding. Originally the possible speed of the car lay in the interval (49,53), which is not entirely contained in the interval (50, ∞) which represents speeding. The second interval, however, is (50.5, 52.5); this *is* contained in (50, ∞). We can represent the policeman's *later* situation as symbolized by $K(\textit{Speeding})$, and the earlier one as $\Diamond K(\textit{Speeding}) \wedge L\Diamond K\neg(\textit{Speeding})$. In the earlier case the motorist *was* speeding, and the policeman had the possibility of knowing this, but did not actually know it.

4.1 Preliminary examples

At this point, we present two simple spaces where we can interpret the language, and some formulas. These should help the reader to become familiar with the semantics.

First, consider the case when X is the set R of real numbers, and $\mathcal{O}$ is the standard topology on R. Suppose that there are two atomic predicates P and I, and that $i(P) = [0, 2]$, and $i[I] = \{x \in R : x \text{ is irrational}\}$. Then $(1, (0, 3)) \models P$, Also, since $2.5 \in (0, 3)$, $(1, (0, 3)) \models L\neg P$. Moreover, $(1, (0, 3)) \models \Diamond KP$, since we can shrink $(0, 3)$ around 1 to $(.5, 1.1)$, say, and have the new neighborhood entirely inside the interpretation of P. On the other hand, $(2, (0, 3)) \not\models \Diamond KP$. The reason is that every open u around 2 contains a point larger than 2, and so $(2, u) \models \neg KP$. We also write this last fact as $(2, (0, 3)) \models \Box L\neg P$. For a final example, $(0, R) \models K\Box LI$, since every open set containing any real is non-empty and so contains an irrational.

The next example itself will re-appear (in a slightly elaborated manner) in Sec. 6 as Example A. A picture of it may be found in Fig. 6.3 below.

Let $X = \{a, p, q, z_1, z_2\}$, and let $\mathcal{O}$ contain X and

$$\begin{array}{ll} u_1 = \{a, p, z_1\} & u_2 = \{a, q, z_2\} \\ v_1 = \{a, z_1\} & v_2 = \{a, z_2\}. \end{array}$$

Let P and Q be atomic sentences; we form a subset space $\langle X, \mathcal{O}, i\rangle$ via

$$i(P) = \{p\} \qquad i(Q) = \{q\}$$

Note that $a, v_1 \models K\neg P$. the reason for this is that neither a nor z_1 satisfy P. Since $v_1 \subseteq u_1$, we see that $a, u_1 \models \Diamond K\neg P$. So we have

$$a, u_1 \models (LP \wedge \neg LQ) \wedge \Diamond K\neg P.$$

Similarly, we have

$$a, u_2 \models (LQ \wedge \neg LP) \wedge \Diamond K\neg Q.$$

We conclude that

$$a, X \models \Diamond((LP \wedge \neg LQ) \wedge \Diamond K\neg P) \ \wedge \ \Diamond((LQ \wedge \neg LP) \wedge \Diamond K\neg Q). \tag{6.3}$$

4.2 Further definitions

Certain kinds of sentences will have special interest in our study. Given a model $\mathcal{X}$, and a sentence ϕ, ϕ is *persistent in* $\mathcal{X}$ if for all p, u, v so that $p \in v \subseteq u$, we have that if $p, u \models_{\mathcal{X}} \phi$ then $p, v \models_{\mathcal{X}} \phi$. ϕ is *persistent* if it is persistent in all $\mathcal{X}$. ϕ is *bi-persistent* if for all $\mathcal{X}$ and all p, u, v so that $p \in v \subseteq u$, we have $p, u \models_{\mathcal{X}} \phi$ iff $p, v \models_{\mathcal{X}} \phi$. A sentence ϕ is *reliable* in a model $\mathcal{X}$ if $K\phi \to K\Box\phi$ is valid in $\mathcal{X}$. ϕ is *reliable* if it is reliable in every $\mathcal{X}$. In other words, once ϕ is known, we need not worry about its becoming false. A sentence of the form $K\Box\phi$ is itself always reliable. Reliable sentences represent reliable knowledge and have a rather intuitionistic flavor. However, our logic is classical, since we

are trying to represent certain knowledge theoretic ideas in a classical setting, rather than use an intuitionistic setting where such ideas would be *presupposed*. If the topology is discrete, then the only reliable sentences will be persistent. In comparison, with the trivial topology, only *tautologies* will tend to be reliable in $\mathcal{X}$. Thus, for example, assuming that all boolean combinations of $i(A)$ and $i(B)$ are non-empty, then the only sentences involving A and B which are reliable will be tautologies. Note that when v is a subset of u, then every reliable sentence known at p, u is also known at p, v. This is in accord with the intuition that refining from u to v increases knowledge.

Our language allows us to express certain basic topological notions. If $\mathcal{X}$ is indeed a topology, then a set $i(A)$ will be open iff every point in $i(A)$ has an open neighborhood contained entirely in $i(A)$ iff at any p in $i(A)$, the sentence $\Diamond KA$ holds. Thus $i(A)$ is *open* iff the sentence $A \to \Diamond KA$ is valid in the model. Dually, $i(A)$ is *closed* iff the sentence $\Box LA \to A$ is valid in the model. It is not hard to see that with the obvious definitions, r.e. subsets of the natural numbers will satisfy the same knowledge theoretic sentences that opens do in a topological setting, and this, we believe, is the source of our intuition that there is similarity between open sets and r.e. sets. The set $i(A)$ is *dense* iff the sentence $\Box LA$ is valid and it is *nowhere dense* if $\Diamond L\neg A$ is valid.

4.3 Topologic and the Tarski-McKinsey semantics

We now relate the original topological interpretation of the modal logic $\mathcal{L}_0$ from Sec. 3 to the bimodal logic $\mathcal{L}$ that we have been discussing. Define a map

$$^* : \mathcal{L}_0 \to \mathcal{L}$$

by recursion:

$$\begin{array}{lcl} A^* & = & A \\ (\phi \wedge \psi)^* & = & \phi^* \wedge \psi^* \\ (\neg\phi)^* & = & \neg(\phi^*) \\ (I\phi)^* & = & \Diamond K\phi^*. \end{array}$$

Recall that we interpret $\mathcal{L}$ on all subset models, hence on all topological models. It is on this class that it makes sense to compare the two languages.

PROPOSITION 6.7 *For all ϕ in $\mathcal{L}_0$, all topological models $\mathcal{X}$, and all points $x \in X$,*

$$x \models \phi \quad \textit{iff} \quad x, X \models \phi^* .$$

The proof uses the fact that ϕ^* is always bi-persistent.

The point here is that the original language $\mathcal{L}_0$ corresponds to a *fragment* of the bimodal language $\mathcal{L}$. It should not be surprising that the larger language is strictly more expressive.

Substitution instances of classical tautologies

$K(\phi \to \psi) \to (K\phi \to K\psi)$ $\quad$ $\Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$
$K\phi \to (\phi \wedge KK\phi)$ $\quad$ $\Box\phi \to (\phi \wedge \Box\Box\phi)$
$L\phi \to KL\phi$

$(A \to \Box A) \wedge (\neg A \to \Box\neg A)$ for atomic A

$K\Box\phi \to \Box K\phi$ (the Cross Axiom)

$$\frac{\phi \to \psi,\ \phi}{\psi} \qquad \frac{\phi}{K\phi} \qquad \frac{\phi}{\Box\phi}$$

Figure 6.1. The axioms and rules of the logic of subset spaces.

PROPOSITION 6.8 *The sentence Lp is not equivalent on the class of topological spaces to ϕ^* for any sentence ϕ of $\mathcal{L}_0$. That is, there is no $\mathcal{L}_0$-sentence ϕ such that for all topological spaces $\mathcal{X}$, all $x \in X$, and all interpretations i of atomic sentences in X,*

$$x \models \phi \quad \textit{iff} \quad x, X \models Lp.$$

Proof Here is a sketch. Consider two models, both with the same universe R of reals. In M_1, $i(p) = \emptyset$, and in M_2, $i(p) = \{1\}$. (M_1 was presented in Example 6.1.) An induction on $\psi \in \mathcal{L}_0$ shows that the interpretation of ψ in the two models is the same except possibly for the point 1. In particular, $0 \models_1 \psi$ iff $0 \models_2 \psi$. Now suppose that ϕ exists as in our proposition. Note that $0 \models_1 \neg Lp$ and $0 \models_2 Lp$. It follows that $0 \models_1 \neg\phi$ and $0 \models_2 \phi$; this is a contradiction. QED

5. A logical system: the subset space axioms

One main technical goals of this chapter is to present axiomatizations of the validities of several classes of subset space models: all spaces, intersection spaces, lattice spaces, and complete lattice spaces. The *logic of subset spaces* is described by axioms and rules of inference in Fig. 6.1.

These axioms and rules say that K is $S5$-like, and $\Box$ is $S4$-like. We have an axiom of *atomic permanence*: $(A \to \Box A) \wedge (\neg A \to \Box\neg A)$. This is only sound for atomic A. The intuition is that since an atomic sentence A is true at a point irrespective of which open we are considering, shrinking the open does not alter the truth value of A.

Perhaps the characteristic axiom of the system is $K\Box\phi \to \Box K\phi$, called the *Cross Axiom*. It is the one axiom relating the two modalities. it is often used in its dual form $\Diamond L\phi \to L\Diamond\phi$. Let us check its soundness in this form. Fix a subset model $\mathcal{X}$, and assume that $p \in u \in \mathcal{O}$ has the property that $p, u \models \Diamond L\phi$. We must show that $p, u \models L\Diamond\phi$. Our assumption first gives an open $v \in \mathcal{O}$

such that $p \in v \subseteq u$ and $p, v \models L\phi$. From this, there is some $q \in v$ such that $q, v \models \phi$. Now $q \in u$ as well, and we use this point to show that indeed $p, u \models L\Diamond\phi$. That is, we claim that $q, u \models \Diamond\phi$. The reason again is that $q \in v \subseteq u$, and we have seen above that $q, v \models \phi$.

The axioms and rules of inference are sound for subset spaces. The argument is routine, and perhaps the only interesting part concerns the Cross Axiom; we saw it just above. We defer discussion of completeness to Sec. 7 and decidability to Sec. 8.

5.1 Axioms for spaces with closure properties: the topologic axioms

We next consider what happens in set spaces with various closure properties. We list the axioms of interest in Fig. 6.2.

A *directed space* is one where for every p, u, v with $p \in u$ and $p \in v$ there is a $w \in \mathcal{O}$ such that $p \in w$, $w \subseteq (u \cap v)$ An *intersection space* is one where we can take $w = u \cap v$.

We check that the axiom the Weak-Directedness Axiom (WD) from Fig. 6.2. is sound for directed spaces. Suppose that $x, u \models \Diamond\Box\phi$. Let $v \subseteq u$ be such that $x, v \models \Box\phi$. To see that $x, u \models \Box\Diamond\phi$, let $u' \subseteq u$. Let $w \in \mathcal{O}$ be such that $p \in w \subseteq u' \cap v$. Then since $w \subseteq v$ we have $x, w \models \phi$. Hence $x, u' \models \Diamond\phi$. Since u' is arbitrary, x, u indeed satisfies $\Box\Diamond\phi$.

As it happens, (WD) does not lead to a complete axiomatization of the valid sentences on intersection spaces. We discuss the incompleteness further in Sec. 6.

We next consider the Union Axioms (Un). To check soundness on spaces closed under unions, suppose that x, u^* satisfies the antecedent via u, y, and v such that $u \subseteq u^*$, $y \in u$, $y \in v \subseteq u^*$, $x, u \models \phi$, and $y, v \models \psi$. Let $w = u \cup v$. Then $w \subseteq u^*$. Clearly $x, w \models \Diamond\phi \wedge L\Diamond\psi$. Since each point of w is either in u or v, every point in w has a neighborhood in which either $\Diamond\phi$ or $\Diamond\psi$ is satisfied. The reader might like to consider the scheme (Weak Un) and to see why it actually is weaker than (Un).

The system whose axioms are the basic axioms together with the Weak-Directedness axiom and the Union Axiom will be called topologic. The idea is that topologic should be strong enough to support elementary topological reasoning.

THEOREM 6.9 (GEORGATOS, 1993; GEORGATOS, 1994A)
The topologic *axioms are complete for topological spaces, indeed for complete lattice spaces. Moreover, any sentence satisfiable in any topological space is satisfied in a finite topological space.*

One way to obtain the completeness result goes by considering a canonical model, building on what we have seen in Sec. 3 on the more standard modal

	formal statement + comments
WD	$\Diamond\Box\phi \to \Box\Diamond\phi$
	sound for weakly directed spaces
Un	$\Diamond\phi \wedge L\Diamond\psi \;\to\; \Diamond[\Diamond\phi \wedge L\Diamond\psi \wedge K\Diamond L(\phi \vee \psi)]$
	sound for spaces closed under binary unions
topologic	This is the set space axioms + (WD) + (Un).
	sound for lattice spaces, complete for
	topological spaces – even for complete lattice spaces
Weak Un	$L\Diamond\phi \wedge L\Diamond\psi \;\to\; L\Diamond[L\Diamond\phi \wedge L\Diamond\psi \wedge K\Diamond L(\phi \vee \psi)]$
	weaker than (Un)
CI	$\Box\Diamond\phi \to \Diamond\Box\phi$
	sound for set spaces closed under all intersections
	follows from topologic axioms
M_n	$(\Box L\Diamond\phi \wedge \Diamond K\psi_1 \wedge \cdots \wedge \Diamond K\psi_n)$
	$\to L(\Diamond\phi \wedge \Diamond K\psi_1 \wedge \cdots \wedge \Diamond K\psi_n)$
	(WD) + all (M_n) is complete for directed spaces

Figure 6.2. Axiom schemes for set spaces with additional closure properties.

logic of topology. (See Dabrowski et al., 1996.) This proof also shows that the topologic axioms give a complete axiomatization of the validities on the smaller class of complete lattice spaces. Indeed, the canonical model of topologic turns out to be a complete lattice. Further, it turns out that the (Un) cannot be replaced by (Weak Un) in this axiomatization.

This work also has some model-theoretic corollaries which might be of interest. Here are two of them, again from Dabrowski et al., 1996. In the first, we recall that $\mathcal{L}_0$ is the modal logic of the interior operation I, and the map $^*\ :\ \mathcal{L}_0 \to \mathcal{L}$ is the embedding of $\mathcal{L}_0$ into the bimodal logic of interest here (this map was discussed in Sec. 4.3). We extend the map to sets by $S^* = \{\phi^* : \phi \in S\}$.

THEOREM 6.10 *For all S and ϕ in $\mathcal{L}_0$, the following are equivalent:*

1 $S \vdash \phi$ *in* $S4$.

2 $S^* \vdash \phi^*$ *in* topologic.

In the second result, let Π be the smallest set of sentences containing the atomic sentences, and closed under boolean operations and the operator $\Diamond K$. This is also the image of the *-translation mentioned above. It is easy to check

that each $\pi \in \Pi$ is bi-persistent on all spaces closed under intersections; this can be proven formally using (WD).

THEOREM 6.11 *If ϕ is bi-persistent on the class of finite lattice spaces, then there is some $\pi \in \Pi$ so that $\vdash \psi \leftrightarrow \pi$ in* topologic.

Georgatos also showed that topologic has the finite model property and is therefore decidable. The logic of subset spaces and the logic of intersection spaces do not have the finite model property. However, in Sec. 8 we adapt filtration on certain non-standard models to show that the logic of subset spaces is decidable.

Continuing, we consider the axioms (CI). These are the converse of the (WD) axioms. We claim first that (CI) is sound for complete intersection spaces. To see this, let x be a point in some space $\mathcal{X}$, and assume that $x, v \models \Box\Diamond\phi$. Let u_x be the intersection of all opens containing x. Since u_x has no open proper subsets $x, u_x \models (\Box\phi \vee \Box\neg\phi)$. Hence $x, u_x \models \Box\phi$. Since $u_x \subseteq v$, we have $x, v \models \Diamond\Box\phi$. On the other hand, (CI) is not sound for intersection spaces; this can be shown semantically using Example B of Sec. 6 below. Intuitively, closure under finite intersection alone is not enough to guarantee that the neighborhoods of a point eventually stabilize on ϕ or on $\neg\phi$. The interesting point is that this axiom *is* sound for lattice spaces. So by completeness, each instance is provable in topologic. We know of no purely syntactic argument for this.

We end with a mention of the results on intersection spaces and directed spaces. As we noted before, the (WD) axioms are sound for intersection spaces, even for directed spaces. Since these classes are of topological interest, the reader of this handbook article might find them of interest. It turns out that (WD) is not complete for either class. As Weiss has shown in an unpublished note, the two classes of spaces have the same logic, and a characterization of either class by a set of axioms also leads to a characterization of the other. That turns out to require infinitely many axiom schemes which cannot be reduced to a finite set (see Weiss and Parikh, 2002). This can be achieved using the (M_n) schemes; we omit the proof of their soundness for directed spaces and merely state the relevant result.

THEOREM 6.12 (WEISS AND PARIKH, 2002) *The set space axioms plus the axiom (WD) and the axioms (M_n) are complete for directed spaces.*

The topics described in this section do not exhaust what is known in the field. For example, Georgatos, 1997 studies tree-like spaces. (Given two open sets in a tree-like topological space, they are either disjoint, or one is a subset of the other.) See also Sec. 9 for other results which use languages that are variants of the one presented in Sec. 4 and studied in this section.

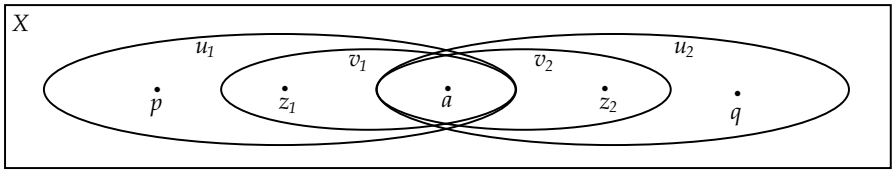


Figure 6.3. Example A.

6. Further examples

In this section, we present two further examples of set spaces and sentences. (The first of these was essentially introduced in Sec 4.1.) These examples are somewhat pathological, and we present them to justify some of the claims in the last section, and also to motivate some of the work in Sections 7 and 8.

Example A. Let $X = \{a, p, q, z_1, z_2\}$, and let $\mathcal{O}$ contain X and

$$\begin{array}{ll} u_1 = \{a, p, z_1\} & u_2 = \{a, q, z_2\} \\ v_1 = \{a, z_1\} & v_2 = \{a, z_2\}. \end{array}$$

Let P, Q, and Z be atomic sentences. (Our work in Sec. 4.1 did not need Z.) We form a subset space $\langle X, \mathcal{O}, i\rangle$ via

$$i(P) = \{p\} \qquad i(Q) = \{q\} \qquad i(Z) = \{z_1, z_2\}.$$

(See Fig. 6.3.)

Our first observation is that the Weak Directedness axioms are validated in this model (despite the fact that the model is not actually closed under intersections). This is an instance of a very general fact. In a subset frame with only finitely many opens, every open u about a point p can be shrunk to a minimal open v about p. When v is minimal (p, v) automatically satisfies all sentences of the form $\phi \to \Box\phi$. In this way, all finite spaces satisfy the Weak Directedness axioms.

Next, we see that the Weak Union axioms also hold (and again, the model is not actually closed under unions). Since there are a number of cases, we only present a few of them. Suppose that $p, u_1 \models \phi$ and $q, u_2 \models \psi$. Then for any $x \in X$ we have $x, X \models L\Diamond\chi$, where

$$\chi \quad \equiv \quad L\Diamond\phi \ \wedge \ L\Diamond\psi \ \wedge \ K\Diamond L(\phi \vee \psi)\ .$$

The most interesting case is where, e.g., $p, u_1 \models \phi$ and $z_2, v_2 \models \psi$. Here, z_1, v_1 satisfies ψ as well, since v_1 and v_2 are isomorphic. Thus for any $y \in u_1$ we have $y, u_1 \models \chi$.

However, the stronger Union Axiom fails; for example

$$\begin{aligned} & z_2, X \models \Diamond K(\neg Q) \ \wedge \ L\Diamond(LP \wedge K\neg Q) \\ \text{but} \quad & z_2, X \not\models \Diamond[LP \wedge K\neg Q]. \end{aligned}$$

Note that, prefixed by an L, this sentence would be satisfied. This was the key idea for showing that the Weak Union Axiom is satisfied.

To summarize: this space satisfies the Weak Union Axiom but not the Union Axiom. Hence the former scheme is properly stronger. There is a second reason for introducing this example, having to do with the theories realizable in various spaces.

This example can also be used to show that the subset space logic and Weak Directedness Axioms are incomplete for the class of intersection spaces.

Concerning unions, we have seen in Sec. 4.1 that

$$a, X \models \Diamond((LP \wedge \neg LQ) \wedge \Diamond K\neg P) \ \wedge \ \Diamond((LQ \wedge \neg LP) \wedge \Diamond K\neg Q). \tag{6.4}$$

The witnesses are u_1 and u_2. Also

$$a, X \models K(Z \ \rightarrow \ \neg(\Diamond(LP \wedge \neg LQ) \ \wedge \ \Diamond(LQ \wedge \neg LP))). \tag{6.5}$$

Let ψ be the conjunction of the sentences in (6.4) and (6.5). We claim that ψ has no models which are closed under unions. For suppose $\mathcal{X}'$ were such a model. Let u_1' and v_1' be witnesses in $\mathcal{X}'$ to (6.4); let u_2' and v_2' be subsets witnessing $a, u_1' \models \Diamond K\neg P$ and $a, v_1' \models \Diamond K\neg Q$ respectively. Let $w' = u_1' \cup v_2'$ (we could also use $u_2' \cup v_1'$). Let $z' \in v_2'$ be a Z-point. Then $z', w' \models LP \wedge \neg LQ$ while $z', u_2' \models LQ \wedge \neg LP$. So

$$z', X \models Z \ \wedge \ \Diamond(LP \wedge \neg LQ) \ \wedge \ \Diamond(LQ \wedge \neg LP).$$

This contradicts (6.5).

This fact that the model satisfies the Weak Directedness Axioms and Weak Union Axioms shows that the basic axioms of topologic together with all of these axioms cannot refute $\psi \wedge \chi$. Nevertheless $\psi \wedge \chi$ cannot hold in any lattice space (even in any directed space). So the axioms are incomplete. (Nevertheless, we do have completeness for lattice spaces using the stronger Union Axioms.)

Next, we present an example which shows a number of things about the theories realized in spaces closed under finite intersections.

Example B. The space $\mathcal{X}$ has points

$$a_0, a_1, a_2, \ldots, a_n, \ldots, \qquad b_0, b_1, b_2, \ldots, b_n, \ldots, \qquad \text{and } c\,.$$

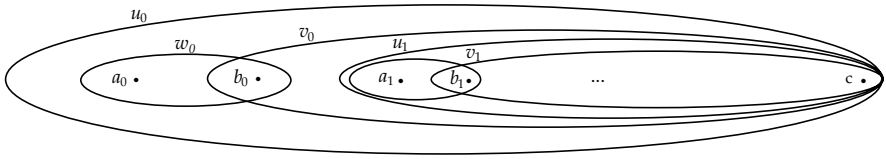


Figure 6.4. Example B.

There are several families of opens

$$\begin{array}{lcl} u_n & = & \{c\} \cup \{a_m : m \geq n\} \cup \{b_m : m \geq n\} \\ v_n & = & \{c\} \cup \{a_m : m > n\} \cup \{b_m : m \geq n\} \\ w_n & = & \{a_n, b_n\} \\ w'_n & = & \{b_n\}. \end{array}$$

(See Fig. 6.4.) We interpret three predicates A, B, and C in the obvious way.

Let

$$\mathsf{last\text{-}and\text{-}}B \quad \equiv \quad B \wedge \Box(KB \vee LC).$$

Informally, $\mathsf{last\text{-}and\text{-}}B$ should hold only at a pair (d, s) only when d is a b-point, and only when d is the last element of s (the element with lowest subscript). Note that $b_i, v_i \models \mathsf{last\text{-}and\text{-}}B$, since every subset of v_i which contains b_i is either the singleton $\{b_i\}$ or contains c. Further, $b_i, u_i \models \neg\mathsf{last\text{-}and\text{-}}B$, since $w_i \subseteq v_j$, and $b_i, w_i \models \neg(KB \vee LC)$. Finally, if $i > j$, then $b_i, v_j \models \neg\mathsf{last\text{-}and\text{-}}B$ for the same reason. So in this space $\mathsf{last\text{-}and\text{-}}B$ has the meaning that we have described.

Another fact about this example is that if $i > j$ and $i' > j'$, then $th(a_i, u_j) = th(a_{i'}, u_{j'})$, and $th(a_i, v_j) = th(a_{i'}, v_{j'})$. Similar statements hold for the b-points, and also $th(c, u_i) = th(c, u_{i'})$, etc. All of these facts are proved by induction. Alternatively, one may use the appropriate version of Fraisse-Ehrenfeucht games for this semantics.

This example also has important consequences for theories in the set space logic. Consider the theories $T = th(c, u_i)$ and $U = th(c, v_i)$. These theories are not equal, since $L(\mathsf{last\text{-}and\text{-}}B)$ is in the latter but not the former. Nevertheless, the two theories have the property that for every $\phi \in T$, $\Diamond\phi \in U$. In Sec. 7, we will introduce the notation $T \xrightarrow{\Diamond} U$ for this relation. In this example, we also see that $U \xrightarrow{\Diamond} T$. Now our intuition about the $\xrightarrow{\Diamond}$ relation is that $T \xrightarrow{\Diamond} U$ *should* mean that in every space containing (d, s) with theory T, it is possible to shrink s to $s' \subseteq s$ with $th(d, s') = U$.

A further desirable result would be that if $T \xrightarrow{\Diamond} U \xrightarrow{\Diamond} T$, then $T = U$. However, the example under consideration shows that this is not in general correct. This accounts for some of the difficulties in the completeness proof for the logic of set spaces.

A related fact is that the set space logic does not have the finite model property. To see this, consider $\phi \equiv L(\mathsf{last\text{-}and\text{-}}B)$, and note that both T and U contain

$$\psi \quad \equiv \quad (\Box\Diamond\phi) \wedge (\Box\Diamond\neg\phi). \tag{6.6}$$

We claim that no sentence of the form (6.6) can have a finite model. For suppose that $\mathcal{X}$ were a finite space containing x and u such that $x, u \models \psi$. We may assume that u is a $\subseteq$-minimal open about x with this property. But the minimality implies that $x, u \models \phi \wedge \neg\phi$, and this is absurd.

7. Completeness of the subset space axioms

The following is the main result on the subset space logic.

THEOREM 6.13 *The basic axioms are strongly complete for subset space models. That is, if $T \models \phi$, then $T \vdash \phi$.*

This section is an extended sketch of the proof, presenting many of the details and ideas but certainly leaving out a good deal of the work. The definitions at the beginning will be used in later sections, as will Propositions 6.14 and 6.15.

As in Sec. 3, we use *theories*; these are the maximal consistent subsets of the language $\mathcal{L}$. Let th be the set of theories in $\mathcal{L}$ using the subset space axioms from Sec. 5. We continue to use letters like T, U, V, etc., to denote theories. In order to prove that we have given a complete proof system, we need only show that for every theory T, there is a subset space model $\mathcal{X} = \langle X, \mathcal{O}, i\rangle$, a point $x \in X$, and a subset $u \in \mathcal{O}$ such that $p, u \models_{\mathcal{X}} T$.

To get started, we define the relations $\xrightarrow{L}$ and $\xrightarrow{\Diamond}$ on theories by:

$$U \xrightarrow{L} V \text{ iff whenever } \phi \in V,\ L\phi \in U$$
$$U \xrightarrow{\Diamond} V \text{ iff whenever } \phi \in V,\ \Diamond\phi \in U.$$

Of course, the maximal consistency of theories give other characterizations. For example, $U \xrightarrow{L} V$ if whenever $K\phi \in U$, $\phi \in V$.

Further, define $U \xrightarrow{L\Diamond} V$ if for all $\phi \in V$, $L\Diamond\phi \in U$. And define relations such as $U \xrightarrow{\Diamond L} V$ and $U \xrightarrow{\Box\Diamond} V$ similarly.

PROPOSITION 6.14 *Concerning the relations $\xrightarrow{L}$ and $\xrightarrow{\Diamond}$:*

(1) $\xrightarrow{L}$ is an equivalence relation.

(2) $\xrightarrow{\Diamond}$ is reflexive and transitive.

(3) If $L\phi \in T$, then there is some U so that $\phi \in U$ and $T \xrightarrow{L} U$.

(4) If $\Diamond\phi \in T$, then there is some U so that $\phi \in U$ and $T \xrightarrow{\Diamond} U$.

Proof These are all standard consequences of the S4-ness of $\Diamond$ and the S5-ness of L. QED

These facts will be used in the sequel without mention. In addition, we have the following consequence of the Cross Axiom.

PROPOSITION 6.15 *Let U and V be theories, and suppose that there is a theory W such that $U \xrightarrow{\Diamond} W \xrightarrow{L} V$.*

$$\begin{tikzcd} U \arrow[r, "\Diamond"] \arrow[d, dashed, "L"'] & W \arrow[d, "L"] \\ T \arrow[r, dashed, "\Diamond"'] & V \end{tikzcd}$$

Then there is a theory T so that $U \xrightarrow{L} T \xrightarrow{\Diamond} V$.

Proof Let $S = \{\Diamond\phi : \phi \in V\} \cup \{\psi : K\psi \in U\}$. We claim that S is consistent; suppose towards a contradiction that it is not. Then there is a finite subset of S which is inconsistent. Now the two sets whose union is S are closed under conjunction, and moreover, K commutes with conjunction. So there are *individual* ϕ and ψ such that $\phi \in V$, $K\psi \in U$ and $\vdash \Diamond\phi \to \neg\psi$. Therefore $\vdash L\Diamond\phi \to L\neg\psi$; hence this sentence belongs to U. And also, $\phi \in V$, so $L\phi \in W$ and $\Diamond L\phi \in U$. So $L\Diamond\phi \in U$ by the Cross Axiom. It follows that $L\neg\psi \in U$. But since $K\psi \in U$, this gives the contradiction that U is inconsistent. So S is consistent. Let $T \supseteq S$ be maximal consistent. By construction, $U \xrightarrow{L} T \xrightarrow{\Diamond} V$. QED

Proposition 6.15 is the embodiment of the Cross Axiom in the realm of theories. The next result is a generalization which will be used in the proof of completeness.

PROPOSITION 6.16 *Let $\mathcal{L} = \langle L, \leq, \top\rangle$ be a finite, bounded, linear order with top element $\top$, and let T be an order preserving map from $\mathcal{L}$ to $\langle th, \xrightarrow{\Diamond}\rangle$. Suppose that $T_\top \xrightarrow{L} U^*$. Then there is an order preserving $U : \mathcal{L} \to th$ such that $U_\top = U^*$, and for all $m \in L$, $T_m \xrightarrow{L} U_m$.*

Proof U is defined going down $\mathcal{L}$, using Proposition 6.15. QED

The first idea in proving Theorem 6.13 is to consider the canonical model and to show that every theory T is the theory of some pair p, u from that model.

As we have seen, the collection th of theories has relations $\xrightarrow{L}$ and $\xrightarrow{\Diamond}$ with the cross property of Proposition 6.15. We shall call such a structure a *cross axiom frame*; we can soundly interpret $\mathcal{L}$ on it using the two relations. (We shall pursue this in Sec. 8.) The canonical cross axiom model shows that the subset space logic is complete for interpretations of $\mathcal{L}$ in cross axiom models. However, the canonical cross axiom model is not a subset space, and there does not seem to be any straightforward way to turn it into one.

Nevertheless, we would like completeness relative to subset spaces. Again, the natural idea is to try to "spatialize" the space of canonical theories (perhaps by taking as opens the L-equivalence classes). However, this construction could not work for the following reasons: Call a model $\mathcal{X}$ *exact* if for every T there are unique p and u so that $th_{\mathcal{X}}(p,u) = T$, and if $u \supseteq v$ iff $th_{\mathcal{X}}(p,u) \xrightarrow{\Diamond} th_{\mathcal{X}}(p,v)$. But in Example B we saw theories $T \xrightarrow{\Diamond} U \xrightarrow{\Diamond} T$ which are distinct. This implies that there are no exact subset space models.

For this reason, we do not approach completeness via the canonical model. The strategy is to build a space X of "abstract" points. We shall also have opens given in an abstract way, via a poset P and an antitone (i.e., order-reversing) map $i : \mathsf{P} \to \mathcal{P}^*(X)$. The points are abstract since they are not theories. But with each x and each p so that $t \in i(p)$ we shall have a "target" theory $t(x,p)$. The goal of the construction is to arrange that in the overall model, $th(x, i(p)) = t(x,p)$.

We are not going to present all of the details here, only the basic ideas. We hope that the interested reader can fill in the rest, based on the following more concrete plan.

One builds

(1) A set X containing a designated element x_0.

(2) A poset with least element $\langle \mathsf{P}, \leq, \bot \rangle$.

(3) A function $i : \mathsf{P} \to \mathcal{P}^*(X)$ such that $p \leq q$ iff $i(p) \supseteq i(q)$, and $i(\bot) = X$. (That is, a homomorphism from $\langle \mathsf{P}, \leq, \bot \rangle$ to $\langle \mathcal{P}^*(X), \supseteq, X \rangle$.)

(4) A partial function $t : X \times \mathsf{P} \to th$ with the property that $t(x,p)$ is defined iff $x \in i(p)$. Furthermore, we require the following properties for all $p \in \mathsf{P}$, $x \in i(p)$, and ϕ:

(a.1) If $y \in i(p)$, then $t(x,p) \xrightarrow{L} t(y,p)$.

(a.2) If $L\phi \in t(x,p)$, then for some $y \in i(p)$, $\phi \in t(y,p)$.

(b.1) If $q \geq p$, then $t(x,p) \xrightarrow{\Diamond} t(x,q)$.

(b.2) If $\Diamond\phi \in t(x,p)$, then for some $q \geq p$, $\phi \in t(x,q)$.

(c) $t(x_0, \bot) = T$, where T is the theory from above which we aim to model.

Suppose we have X, P, i, and t with these properties. Then we consider the subset space

$$\mathcal{X} \quad = \quad \langle X, \{i(p) : p \in \mathsf{P}\}, i\rangle.$$

where $i(P) = \{x : A \in t(x, \bot)\}$.

LEMMA 6.17 (THE TRUTH LEMMA) *Assume conditions (1)–(4) for X, P, i, and t. Then for all $x \in X$ and all $p \in \mathcal{L}$ such that $x \in i(p)$,*

$$th_{\mathcal{X}}(x, i(p)) \quad = \quad t(x, p).$$

Proof We show induction on ϕ that ϕ belongs to the set on the left iff it belongs to the set on the right. We only give the inductive step for sentences $L\phi$.

Suppose that $x, i(p) \models L\phi$. Then there is some $y \in i(p)$ such that $y, i(p) \models \phi$. By induction hypothesis, $\phi \in t(y, p)$. By property (4a.1), $t(x, p) \xrightarrow{L} t(y, p)$. Therefore $L\phi \in t(x, p)$. On the other hand, if $L\phi \in t(x, p)$, then by property (4a.2) there is some $y \in i(p)$ such that $\phi \in t(y, p)$. By induction hypothesis, $y, i(p) \models \phi$. Therefore $x, i(p) \models L\phi$. This concludes the induction step for L.

The induction step for $\Diamond$ is similar and uses (3), (4b.1), and (4b.2). QED

One builds X, P, i, and t by recursion, in a step-by-step process that goes on for countably many steps. This is not the place to enter into these details. They are fairly straightforward and can be found in Dabrowski et al., 1996.

By the Truth Lemma and property (4c) above, $th_{\mathcal{X}}(x_0, \bot) = T$. So the theory that we started with has a model. In the usual way, this proves the Completeness Theorem.

8. Decidability of the subset space logic

Despite the failure of the finite model property, we prove that the logic of subset spaces is decidable. The proof here is due to Krommes, 2003, simplifying the argument from Dabrowski et al., 1996. It goes by showing that a satisfiable sentence ϕ has a finite *cross axiom* model. This is a kind of pseudo-model for this subject, to be defined shortly. The idea as always is that we move to a bigger class of models than the one we are primarily interested in, and this class will turn out to be better behaved.

DEFINITION 6.18 *A cross axiom frame is a tuple $\langle J, \xrightarrow{L}, \xrightarrow{\Diamond}\rangle$ such that J is a set, $\xrightarrow{L}$ is an equivalence relation on J, $\xrightarrow{\Diamond}$ is a preorder on J, and the following property holds: If $i \xrightarrow{\Diamond} j \xrightarrow{L} k$, then there is some l such that $i \xrightarrow{L} l \xrightarrow{\Diamond} k$. A cross axiom model is a cross axiom frame together with an interpretation i of the atomic symbols of $\mathcal{L}$.*

Note that when we interpret the language on a cross axiom model, we have a *node* on the left side of the turnstile. That is, we write, e.g., $j \models \phi$ since there are no sets involved.

The subset space logic is complete for interpretations in cross axiom models since the latter include subset spaces.

Our main example of a cross axiom model which is not a subset space is the *canonical model* of the subset space logic:

$$\mathcal{C}(ca) \quad = \quad \langle th, \xrightarrow{L}, \xrightarrow{\Diamond} \rangle.$$

(The "ca" stands for "cross axiom".) Proposition 6.15 says that $\mathcal{C}(ca)$ actually is a cross axiom model. The standard truth lemma for this structure shows that for all $T \in th$, $th(T) = T$; that is, the set of sentences satisfied by the point T in $\mathcal{C}(ca)$ is T itself. We shall use a version of filtration to prove a finite model property.

Let ϕ be any sentence which is consistent in the logic of set spaces, and consider the following sets:

$\Sigma^{\neg}$ = all subformulas of ϕ and their negations.

Σ' = all conjunctions and disjunctions of (finite) subsets of $\Sigma^{\neg}$.

Σ'' = all conjunctions and disjunctions of (finite) subsets of Σ'.

$\Sigma^{KL} = \{L\psi : \psi \in \Sigma''\} \cup \{K\psi : \psi \in \Sigma''\}$.

$\Sigma = \Sigma'' \cup \Sigma^{KL}$.

Note that Σ is finite. Note also that modulo propositional logic, Σ'' is closed under the boolean connectives; in effect we are taking disjunctive normal forms. Finally, Γ depends on the original ϕ.

Write $U \equiv V$ if $U \cap \Gamma = V \cap \Gamma$, and let $[U]$ be the equivalence class of U under this relation. (Note that $\equiv$ also depends on ϕ, but we save a little notation by suppressing this.) We define relations $\xrightarrow{L}$ and $\xrightarrow{\Diamond}$ on $[\mathcal{C}(ca)]$ as follows:

$$[S] \xrightarrow{L} [T] \quad \text{iff} \quad \text{there exist } S' \in [S] \text{ and } T' \in [T] \text{ such that } S' \xrightarrow{L} T'.$$

This is the definition used in minimal filtrations. In contrast, we want $\xrightarrow{\Diamond}$ to be transitive, and so we define

$$\begin{aligned} [S] \xrightarrow{\Diamond} [T] \quad \text{iff} \quad & \text{there exist } n \geq 0 \text{ and } S', S_1, \ldots, S_n, S'_n, T' \text{ such that} \\ & S \equiv S' \xrightarrow{\Diamond} S_1 \equiv S'_1 \xrightarrow{\Diamond} \cdots \xrightarrow{\Diamond} S_n \equiv S'_n \xrightarrow{\Diamond} T' \equiv T. \end{aligned}$$

We call the tuple

$$[\mathcal{C}(ca)] = \langle [\mathcal{C}(ca)], \xrightarrow{L}, \xrightarrow{\Diamond} \rangle$$

the *quotient of* $\mathcal{C}(ca)$ *by* $\equiv$. We shall show that this quotient is a cross axiom frame; the main points are the transitivity of $\xrightarrow{L}$ and the cross axiom property.

We turn the structure into a model in the evident way, via the interpretation $i(A) = \{[T] : A \in T \cap \Sigma\}$.

Standard reasoning shows that $[\mathcal{C}(ca)]$ is a filtration of $\mathcal{C}(ca)$. It follows that since the sentence ϕ with which we began is consistent, it has a *finite* cross axiom model. Thus the decidability reduces to the verification of the properties of $[\mathcal{C}(ca)]$ that we mentioned in the last paragraph.

We need several results, starting with an important lemma. For each theory T, let

$$\gamma_T \quad = \quad \bigwedge(T \cap \Gamma).$$

That is, we take the conjunction of the set $T \cap \Gamma$. This is an analog of *atoms*, as we find them in the completeness proof of PDL and other places (see Kozen and Parikh, 1981).

LEMMA 6.19 *Let T and U be theories. If $L\gamma_T \in U$, then whenever $V \equiv U$, we also have $L\gamma_T \in V$.*

Proof Fix U containing $L\gamma_T$, and let $V \equiv U$. We split γ_T into two conjuncts

$$\gamma''_T = \bigwedge(T \cap \Gamma'') \qquad \gamma^{KL}_T = \bigwedge(T \cap \Gamma^{KL}).$$

Again, we have $\gamma_T \equiv \gamma''_T \wedge \gamma^{KL}_T$. Our goal is to show that $L\gamma''_T \in V$.

Let $\psi \in T \cap \Gamma^{KL}$. (Notice at this point that ψ belongs to V.) This ψ is either $L\chi$ or $K\chi$ for some $\chi \in \Gamma''$. The S5 laws of K in subset spaces tell us that $K\chi \leftrightarrow KK\chi$ and $L\chi \leftrightarrow KL\chi$. So $\psi \equiv K\psi$. We therefore see that $K\psi \in V$. This for all $\psi \in T \cap \Gamma^{KL}$ shows that $K\gamma^{KL}_T \in V$.

Note second that since U contains $L\gamma_T$, it also contains $L\gamma''_T$. But this last sentence belongs to $\Gamma^{KL} \subseteq \Gamma$. Because $V \equiv U$, we see that $L\gamma''_T \in V$.

We conclude that V contains $K\gamma^{KL}_T$ and $L\gamma''_T$. Thus it contains $L(\gamma^{KL}_T \wedge \gamma''_T)$. But this is equivalent to $L\gamma_T$. So $L\gamma_T \in V$. QED

LEMMA 6.20 *Suppose that $[S] \xrightarrow{L} [T]$. Then for all $S' \in [S]$ there is some $T' \in [T]$ such that $S' \xrightarrow{L} T'$.*

Proof Fix $S' \in [S]$. Let $S'' \in [S]$ and $T'' \in [T]$ be such that $S'' \xrightarrow{L} T''$. Then $L\gamma_{T''} \in S''$. By Lemma 6.19, $L\gamma_{T''} \in S'$. And thus there is some T' such that $S' \xrightarrow{L} T'$ and $\gamma_{T''} \in T'$. Since $\gamma_{T''} = \gamma_T$, we see that $T' \equiv T$. QED

Lemma 6.20 easily implies the transitivity of $\xrightarrow{L}$ in $[\mathcal{C}(ca)]$. We are thus left with the verification of the cross axiom property. Suppose that $[S] \xrightarrow{\Diamond} [T] \xrightarrow{L} [U]$. We need to find W so that $[S] \xrightarrow{L} [W] \xrightarrow{\Diamond} [U]$, and for this we argue by induction on n in the definition of the relation $[S] \xrightarrow{\Diamond} [T]$. When $n = 0$, we have

$S \equiv T$. In this trivial case we have $[S] = [T]$ and may take $[W] = [S]$. Assume our result for n, and suppose that $[S] \xrightarrow{\Diamond} [T]$ via a chain of length $n+1$, say

$$S \equiv S' \xrightarrow{\Diamond} S_1 \equiv S_1' \xrightarrow{\Diamond} \cdots \xrightarrow{\Diamond} S_n \equiv S_n' \xrightarrow{\Diamond} S_{n+1} \equiv S_{n'+1} \xrightarrow{\Diamond} T' \equiv T.$$

Then $[S_1] \xrightarrow{\Diamond} [T]$ via a chain of length n, and so by induction hypothesis we have some W such that $[S_1] \xrightarrow{L} [W] \xrightarrow{\Diamond} [U]$. By Lemma 6.20 there is some $W' \equiv W$ such that $S_1 \xrightarrow{L} W'$. So as theories, $S' \xrightarrow{\Diamond} S_1 \xrightarrow{L} W'$. Since $\mathcal{C}(ca)$ is a cross axiom frame, there is some X such that $S' \xrightarrow{L} X \xrightarrow{\Diamond} W'$. We have $[S] \xrightarrow{L} [X]$ in the quotient. We also have $[X] \xrightarrow{\Diamond} [W] \xrightarrow{\Diamond} [U]$. By transitivity we also have $[X] \xrightarrow{\Diamond} [U]$. This completes the proof of decidability: from a consistent sentence, we have effectively produced a finite model.

9. Heinemann's extensions to topologic

This section discusses two of the many extensions of topologic due to Bernhard Heinemann. One of his papers (Heinemann, 1998) studies spaces which satisfy *chain conditions* of various sorts. We begin with the relevant definitions; let $\mathcal{X} = \langle X, \mathcal{O} \rangle$ be a subset frame.

1 $\mathcal{X}$ satisfies the *weak bounded chain condition* (*wbcc*) if for each point $x \in X$, every descending sequence of opens around x is finite.

2 $\mathcal{X}$ satisfies the *finite chain condition* (*fcc*) if every descending sequence of opens is finite.

3 $\mathcal{X}$ satisfies the *bounded chain condition* (*bcc*) if for some n, every descending sequence of opens is of length at most n.

In order to axiomatize the logics of these classes of spaces, it is necessary to alter the basic semantics of topologic. Up until now we have

$$p, u \models \Box\phi \qquad \text{iff} \qquad p, v \models \phi \text{ for all } v \in \mathcal{O} \text{ such that } v \subseteq u.$$

In the study of chain conditions, we alter this to

$$p, u \models \Box\phi \qquad \text{iff} \qquad p, v \models \phi \text{ for all } v \in \mathcal{O} \text{ such that } v \subset u.$$

(So $\Box$ now quantifies over *proper* subsets.) Turning to the logics themselves, first note that the topologic axiom $\Box\phi \to \phi$ is no longer valid. So this axiom is dropped, and the rest of the system is retained.

To axiomatize the validities in the wbcc spaces, one then adds the scheme

$$\Box(\Box\phi \to \phi) \to \Box\phi. \tag{6.7}$$

This scheme is familiar from provability logic (also called Löb logic). The resulting logic lacks the finite model property, and indeed the decidability remains open (see Heinemann, 1998).

Turning to bcc spaces, we would add (6.7) (actually a weaker scheme suffices) and also

$$K\Box(K\Box\phi \to \phi) \to K\Box\phi).$$

It turns out that the resulting system is sound for fcc spaces and complete even for the smaller class of bcc spaces, and that it is decidable.

Around the same time as we wrote this chapter, Heinemann discovered a modality which is basic for our enterprise. This is *overlap operator* O, (see Heinemann, 2006). We *add* it to the basic language of topologic and then the semantics is extended to

$$p, u \models \mathsf{O}\phi \qquad \text{iff} \qquad p, v \models \phi \text{ for all } v \in \mathcal{O} \text{ such that } p \in v.$$

Note that the difference between this and the semantics of $\Box\phi$ is that we do not require that $v \subseteq u$. As a result, we see that the axiom $\mathsf{O}\phi \to \Box\phi$ is valid. Also, the S5 axioms are valid for O:

$$\begin{array}{l} \mathsf{O}(\phi \to \psi) \to (\mathsf{O}\phi \to \mathsf{O}\psi) \\ \mathsf{O}\phi \to \phi \wedge \mathsf{OO}\phi \\ \phi \to \mathsf{OP}\phi. \end{array}$$

In the last of these, the operator P is the dual of O, so $\mathsf{P}\phi$ is defined by $\neg\mathsf{O}\neg\phi$. Turning from axioms to rules of inference, we see easily that the necessitation for O preserves validity. So we get a logical system called ET by adding the axioms and rule to the system of topologic. Heinemann proves completeness and decidability of this system (see Heinemann, 2006). The details are elaborations of the arguments which we have already seen for the basic system of topologic, the subset space axioms.

Recall that the class of directed spaces is not finitely axiomatizable (see Weiss and Parikh, 2002). This is remedied in the larger language: consider

$$\mathsf{P}\Box\phi \to \Diamond\phi.$$

Further, the *weak connectedness* of the space may be captured. This condition says that given any two overlapping sets, one is included in the other. The relevant axiom is

$$(\phi \to \Diamond\psi) \vee \mathsf{O}(\psi \to \Diamond\phi).$$

But this is not so topologically natural. It would be more in the spirit of things to capture the condition that two overlapping sets have a common superset. The natural move here is to add *nominals* I, J, etc. whose values are taken to be opens of the underlying space. Then one defines the semantics as in hybrid

logic, using interpretations on the new nominals and the obvious satisfaction relation. The sentence

$$I \wedge \mathsf{P}J \to \mathsf{P}(\Diamond I \wedge \Diamond J)$$

then captures the condition we mentioned. One can also add nominals for points, again following the lead of hybrid logic . There is an outline of the completeness and decidability of the resulting logical system in Heinemann, 2006.

We close with a mention of an avenue for further work. Most of the results in this area concern completeness and decidability of various systems, and also results on expressive power. Results on *complexity* are scarce. One exception is (Heinemann, 1999a), where an NP–completeness result is proved for a satisfiability problem related to *topological nexttime logic*. But again, there are many open complexity problems.

10. Common knowledge in topological settings

The interaction of the topological semantics of modal logic with the topic of *common knowledge* is the topic of a recent paper (van Benthem and Sarenac, 2004). Like those discussed in Sec. 9, we expect this paper to inspire others. Our purpose here is to give a high-level overview of it that connects it with the theme of our chapter.

The most standard setting of modal logic is that given by the semantics on Kripke models. We shall go into more detail on this in Sec. 11 just below. In the standard setting, one adds *transitive closure* operators $\Box^*$ to modal logic. For the semantics, one takes a model M which lives on a relation R, considers the reflexive-transitive closure R^* of R, and finally defines

$$x \models \Box^*\phi \qquad \text{iff} \qquad y \models \phi \text{ for all } y \in M \text{ such that } xR^*y.$$

So $\Box^*\phi$ is semantically equivalent to the infinite conjunction $\phi \wedge \Box\phi \wedge \Box^2\phi \wedge \cdots$. We sometimes read $\Box^*\phi$ as *fully-aware* knowledge; ϕ is true, the agent knows this, s/he knows that s/he knows *this*, etc. But note that this reading hides a great deal. Indeed, it would be better to note that there is an intuitively important idea of fully-aware knowledge and then to take $\Box^*$ as a *proposal* for modeling it. A second proposal is to take the set of states satisfying $\Box^*\phi$ to be the largest fixed point of the following operator on the power set of the set of states in our model:

$$X \mapsto \{w \in W : w \models \phi, \text{ and all } R \text{ successors of } w \text{ belong to } X\}.$$

Yet another proposal might be to have some very explicit evidence, and of course this is not capturable in the standard semantics in the first place. (In the same way, we should always remember that the standard semantics of $\Box$ is a proposal for the modeling of intuitive knowledge in the first place.)

To capture this new operator $\Box^*$ in a sound, complete, and decidable logic, one adds necessitation for the new operator, an axiom called Mix:

$$\Box^*\phi \to (\phi \wedge \Box\Box^*\phi)$$

and also an induction rule: from $\psi \to \phi \wedge \Box\psi$, infer $\psi \to \Box^*\phi$.

So far in our chapter we have only dealt with epistemic logic with one agent, and this is how the subject of epistemic logic germinated. But the whole subject blossoms in the *multi-agent* setting. For simplicity, we shall deal with two agents, call them A and B. The intuitive idea of common knowledge is the two-agent version of the fully-aware knowledge that we saw above. So we add an operator $\Box^*$ (sometimes it is decorated with the names of the agents A and B) to the basic modal language. Then the question arises as to the semantics. One way to begin is to start with a model M living on two different relations, say R_A and R_B. The standard proposal for the semantics of $\Box^*\phi$ is

$$x \models \Box^*\phi \qquad \text{iff} \qquad y \models \phi \text{ for all } y \in M \text{ such that } x(R_A \cup R_B)^*y$$

So we take the reflexive-transitive closure of the union of the two relations. In plainer terms, this proposal amounts to saying that ϕ is common knowledge to A and B if

(0) ϕ is true;

(1) A and B both know (0);

(2) A and B both know (1);

and so on. To summarize, this standard proposal is to identify common knowledge of ϕ with a certain infinite conjunction of iterated knowledge statements of ϕ.

We have taken great care in this discussion to keep separate the intuitive concepts and the standard formulation of them. These are usually conflated, and the main point of van Benthem and Sarenac's paper is that a two-agent version of the topological semantics of epistemic logic allows one to distinguish between different formalizations that are always identified in the relational semantics. The need to do this goes back to an influential paper on common knowledge (Barwise, 1988). In that paper, Barwise wishes to question the basic modeling of common knowledge as an infinite iteration. More precisely, he wishes to distinguish between the infinite iteration and the related fixed point; he also is concerned with the notion of "agents having a shared situation".

Recall the topological semantics of epistemic logic, and especially the translation of the interior operator I from Sec. 4.3. So the analog of $\Box^*\phi$ is now

$$\phi \wedge K\Diamond\phi \wedge K\Diamond K\Diamond\phi \wedge \cdots$$

And the natural operator on the power set of the state set is

$$X \mapsto \{w \in X^{\mathrm{o}} : w \models \phi\}$$

(Note that we are using the interior operator here.) For a sentence in the logic at hand, the two definitions agree. And one of the points of the paper of van Benthem and Sarenac is to show a separation of the two formalizations on one particular model for the two-agent version of the logic. This model is

$$Q \times Q \quad = \quad \langle Q \times Q, \mathcal{O}_1, \mathcal{O}_2 \rangle.$$

Thus the states are pairs of rational numbers. A set X is open in $\mathcal{O}_1$ if for all $(r, s) \in X$ there is some rational ϵ such that

$$\{r\} \times ((s - \epsilon, s + \epsilon) \cap Q) \subseteq X.$$

And $\mathcal{O}_2$ is defined similarly. The analog of the infinite iteration now reflects two agents:

$$\phi \wedge \bigwedge_i \Box_i \phi \wedge \bigwedge_i \bigwedge_j \Box_i \Box_j \phi \wedge \cdots$$

(so that i ranges over $\{1, 2\}$). The fixed point now is for the operator which takes a set X to the intersection of the interiors of X in the two topologies. Again, the paper proves that the two notions differ. See also van Benthem et al., 2007 and Sections 3.2 and 3.4 of Ch. 5 for this and related material, such as the completeness of the two-agent version of S4 on the rational square $Q \times Q$.

11. The topology of belief

We have already seen in Sec. 3 that $S4$, a logic of knowledge, has a topological semantics. For the purposes of this chapter, the important conclusion is that the interior operator acts like *knowledge*. In more detail, the properties of this operator as rendered in modal terms correspond to the axioms of a logic $S4$ that may be considered a *logic of knowledge*. The purpose of this section of our chapter is to show that there is a topological operator which acts like *belief* in this same sense. Our work in this section is based on the recent dissertation by one of the authors (Steinsvold, 2006).

Most commonly, the difference between knowledge and belief is taken to be that belief should not imply truth; this corresponds to the assertion that analogues of statements like $I\phi \to \phi$ should *not* be valid. To emphasize that we are dealing with belief, we change the modality to B in this section. So again, we do not want to work with a logic that includes the T-scheme $B\phi \to \phi$.

The most standard logic of belief is the logic $KD45$. $KD45$ is axiomatized using the schemes listed in Fig. 6.5, together with the rules of Modus Ponens and Necessitation which we have seen in Sec. 3. The D scheme corresponds to

the assertion that beliefs are consistent; an agent who believes ϕ should not also believe $\neg\phi$. The 4 and 5 schemes correspond to assertions of *introspection*: if an agent believes ϕ, then they believe the assertion that they believe ϕ. This is the content of 4. As for 5, it says that if an agent does not believe ϕ, then again they believe the assertion of that disbelief. Despite obvious problems, $KD45$ is a standard logic of belief in the sense that the provable sentences may be taken as a first approximation to the properties of belief. Here is the plan of this section of our chapter. We first recall and compare the relational and topological semantics for $KD45$. We then discuss the *derived set* operation from topology (defined in Sec. 11.2 below). and provide an interpretation of $KD45$ in it. Finally, a topological completeness proof is presented for $KD45$ with respect to a class of spaces that we call DSO spaces.

We should note in passing that other authors have explored the properties of this derived set operation using modal logic; see, e.g., Esakia, 1981; Shehtman, 1990; Bezhanishvili et al., 2005. However, none of these works offered epistemic connections. This is one of our goals. Actually, we are more interested in *doxastic* connections, that is connections to the notion of *belief*.

Before plunging into the details, we should mention why we chose this topic for a chapter on Topology and Epistemic Logic. We have already seen various logics in the paper, including a standard logic of knowledge ($S4$) and a standard logic with common knowledge (the logic of $\Box^*$). One certainly can view what we are doing here as being work in the same general direction: find a logic corresponding to a semantics that is already of interest. But this section can also be read in the other direction: given a logic that we believe to be sensible or at least worthy of study, construct a semantics for it. The work of this section suggests a topological development along these lines. In a nutshell, the proposal is to read belief as the dual of the derived set operator on a topological space satisfying some conditions. For a complete discussion of the philosophical aspects of this proposal, see Steinsvold, 2006.

11.1 Relational semantics of $KD45$

We have already discussed the notion of a *frame* in Sec. 3.1. We quickly moved from frames in general to preorders there, but here we need the more general notion. Again, a *frame*, $F = \langle W, R\rangle$, is a set W with a relation R on W. We'll refer to the members of W as points or worlds, interchangeably. If x bears the relation R to y, we'll write either xRy or $(x, y) \in R$. A *model* is a frame together with an interpretation i of atomic sentences in it. We say that the model is *based* on the frame.

The semantics on models is given in the usual way, with the clause for B that

$$w \models B\phi \quad \text{iff} \quad (\forall z)(wRz \text{ implies } z \models \phi).$$

Ax	formal statement	relational correspondent
K	$B(\phi \to \psi) \to (B\phi \to B\psi)$	(none)
D	$B\phi \to \neg B\neg\phi$	every point has an R-successor
4	$B\phi \to BB\phi$	R is transitive
5	$\neg B\phi \to B\neg B\phi$	R is Eucldiean

Figure 6.5. Axiom schemes of $KD45$ with their relational correspondents.

So B is defined using a universal quantifier, like I was in our earlier work.

We read $w \models \phi$ as saying that ϕ *is true at* w or that w *satisfies* ϕ. A sentence ϕ is *valid in a model* iff ϕ is true at every point in the model. A sentence ϕ is *valid in a frame* iff ϕ is valid in every model based on the frame.

The correspondences of Fig. 6.5 are well known. The precise nature of this correspondence is that a frame F satisfies each instance of D (say) iff F meets the condition that every point in it has a successor. A model which meets the condition corresponding to D (that every point have a successor) is called *serial*. The Euclidean condition corresponding to the 5 axioms is

$$(\forall x)(\forall y)(\forall z)((xRy \wedge xRz) \to yRz).$$

$KD45$ is complete with respect to models which have the properties listed above. That is, if $T \cup \{\phi\}$ is a set of sentences, then $T \vdash \phi$ in $KD45$ iff for all serial, transitive, and Eucldiean models M and all points $x \in M$, if $x \models T$, then $x \models \phi$.

This completeness is a parallel to facts we saw in Sec. 3 for $S4$: One considers the set of theories T in $KD45$ and gets the rest of a "canonical model" structure of a model using (6.2) and (6.1). $KD45$ also is decidable via the finite model property.

11.2 The derivative operation on topological spaces

We are now going to be proposing a *different* semantics for modal logic, one based on an operation from topology called the *derived set* operation. We collect in this section some topological preliminaries about this operation. Let $\mathcal{X} = \langle X, \mathcal{O}\rangle$ be a topological space. For a set $A \subseteq X$, we define the *derived set* of A, $d(A)$,

$$w \in d(A) \text{ iff } (\forall U \in \mathcal{O})(\text{if } w \in U \text{ then } (\exists x \in U \setminus \{w\})(x \in A)).$$

In words, w belongs to the derived set of A iff every open set U around A contains some point of A different from w. It also might be useful to write out the complement:

$$w \notin d(A) \text{ iff } (\exists U \in \mathcal{O}))\ (w \in U \text{ but } (\forall x \in U \setminus \{w\})(x \notin A)).$$

This *derived set* $d(A)$ has many other names in the literature on point-set topology: *derivative*, *Cantor-Bendixson derivative*, *set of limit points*, *set of accumulation points*, and *set of cluster points* (of A). It is usually written as A'.

EXAMPLE 6.21 Let $\mathcal{R}$ be the reals with the usual topology, and let

$$A \quad = \quad \left\{ \frac{1}{n} : n = 1, 2, \ldots \right\}$$

Then $d(A) = \{0\}$. That is, every open set containing 0 contains some number $1/n$. But for all $r \neq 0$, there is an open neighborhood of r which contains no number in A. Next, consider

$$B \quad = \quad \left\{ \frac{1}{n} + \frac{1}{n+1+m} : n, m = 1, 2, \ldots \right\}.$$

This is like A, but with a copy of A itself next to each of its points. The copies are disjoint. Then $d(B) = A \cup \{0\}$. And so $d(d(B)) = \{0\}$. Further, $d(d(d(B))) = \emptyset$. More generally, the derived set of any singleton in any topological space is always empty.

EXAMPLE 6.22 Here is another example of how d works, one which will be elaborated in Sec. 11.4 below. Let W be an infinite set, let D be an infinite subset of W, and let $\mathcal{O}$ be the family of subsets $U \subseteq W$ such that $D \setminus U$ is finite, together with the empty set. It is easy to see that $\mathcal{O}$ is closed under finite intersections, since

$$D \setminus (U \cap V) \quad = \quad (D \setminus U) \cup (D \setminus V),$$

and arbitrary unions. (In fact, every superset of a nonempty open set is itself open.) In this way, $\mathcal{W} = \langle W, \mathcal{O} \rangle$ is a topological space.

We compute $d(A)$ for all $A \subseteq W$:

$$d(A) \quad = \quad \begin{cases} \emptyset & \text{if } D \cap A \text{ is finite} \\ A & \text{if } D \cap A \text{ is infinite} \end{cases}$$

In the first case, $D \setminus (D \setminus A) = D \cap A$ is finite. For $w \in W$, let $U_w = \{w\} \cup (D \setminus A)$. Then U_w is an open set containing w, but for all $x \in U_w \setminus \{w\}$, $x \notin A$. So $w \notin d(A)$. Thuse $d(A) = \emptyset$. We turn to the second case. Let $w \in W$ and let U be an open set containing w. We claim that $U \cap A \cap D$ is infinite. For suppose not. Then $D \cap A$ is infinite but $U \cap (A \cap D)$ is finite. So $(D \cap A) \setminus U$ is infinite. So its superset $D \setminus U$ is also infinite. But by definition of the topology, $D \setminus U$ is finite. This contradiction shows that indeed $U \cap D \cap A$ is infinite. In particular, U contains an element of A different from the original w.

EXAMPLE 6.23 Here is a final example, one suggestive of the connection between the relational semantics and the derived set operation. Let $M = \langle W, R\rangle$ be a set with a transitive relation on it. (We do not require that R be reflexive.) Recall the Alexandrov topology from Sec. 3.1: the opens are the R-closed sets. We again compute d on sets, and this time

$$d(A) \quad = \quad \{x \in W : (\exists y \in A)\ (y \neq x \text{ and } xRy)\}.$$

Here is the reasoning: suppose first that $x \in d(A)$. Consider the open set $U_x = \{x\} \cup \{y : xRy\}$. A must contain some point of U_x other than x. In the other direction, assume that $y \neq x$, xRy, and $y \in A$. Then every open set around x contains y.

A suggestive observation: the standard Kripke semantics of modal logic is *almost* related to the derived set by

$$i(\Diamond\phi) \quad = \quad d(i(\phi)).$$

Indeed, if we modified the definition of $x \in d(A)$ to not require witnesses different from x in the open neighborhoods, we would have the equation above. Or, if R were irreflexive, then the equation would hold. See Esakia, 2001 for more on this.

Proposition 6.24 below contains standard results in point-set topology about the operation d. In the first part, we recall that a topological space satisfies the T_1 separation property iff for all distinct points x and y, there is an open set containing x but not y. Another formulation: all singletons are closed sets. This property can also be neatly expressed using d: $d(\{x\}) = \emptyset$ for all x. (Recall also that $x \notin d(\{x\})$.) In the second part, we mention the T_d separation property. This holds if every singleton is the intersection of an open and a closed set. Equivalently, every set of the form $d(A)$ is closed. This alternate formulation is more useful in this chapter.

PROPOSITION 6.24 *Concerning the derivative operation d on a topological space $\mathcal{X}$:*

1 *If every set of the form $d(\{z\})$ is open, then $\mathcal{X}$ is a T_1 space.*

2 *If $\mathcal{X}$ is a T_1 space, then $\mathcal{X}$ is a T_d space.*

3 *For all topological spaces, $d(d(A)) \subseteq A \cup d(A)$.*

Proof For the first assertion, assume that each $d(\{x\})$ is open. Suppose towards a contradiction that $y \in d(\{x\})$. As we know, $x \notin d(\{x\})$. But since $d(\{x\})$ is open, contains y, and is disjoint from $\{x\}$, we see that $y \notin d(\{x\})$ after all. This is a contradiction.

For the second assertion, consider a derived set $d(A)$. Let $x \notin d(A)$. Then there is an open set U containing x such that $(U \setminus \{x\}) \cap A = \emptyset$. We claim that $U \cap d(A) = \emptyset$. Since we already know $x \notin d(A)$, we only need to consider some $y \neq x$ in U. By the T_1 property, let V be an open set with $y \in V$ but $x \notin V$. Then $(U \cap V) \cap A = \emptyset$. Since $U \cap V$ is open and contains y, we see that $y \notin d(A)$. But y is arbitrary, and this establishes the claim that $U \cap d(A) = \emptyset$. That is, U is an open set, $x \in U$, and $U \subseteq X \setminus d(A)$. Since x is arbitrary, $X \setminus d(A)$ is open.

For the last part, suppose that $x \in d(d(A))$ but $x \notin A$. Let U be an open set containing x; we show that $U \cap (A \setminus \{x\}) \neq \emptyset$. Let $y \in d(A) \cap U$. Since $y \in d(A) \cap U$, let $z \in U \cap (A \setminus \{y\})$. Then $z \neq x$, since $x \notin A$. QED

The assertions in Proposition 6.24 suggest a definition: we say that a space $\mathcal{X}$ is a *DSO space* if every derived set $d(A)$ is open but the space is dense in itself. (This last condition means that there are no open singletons. We add this because the conjunction of these two properties will be important in our later work: the DSO spaces turn out to play the role of serial, transitive, Euclidean relations in the Kripke semantics of $KD45$.)

EXAMPLE 6.25 Every space built as in Example 6.22 is a DSO space. To see this, recall that we have computed the derivative operation on all subsets of the space. The upshot is that the only derived sets are the empty set and the whole space. So every derived set is open. Further, no singletons are open. So we have a DSO space.

EXAMPLE 6.26 On the other hand, consider a space $\langle W, \mathcal{O} \rangle$ built from a transitive relation $\langle W, R \rangle$ as in Example 6.23. An open singleton corresponds to a point x with no R-successors different from x itself. We do not usually have a DSO space: usually there are derived sets which are not open. (Incidentally, there are no finite DSO spaces: such a space would be T_1, and hence all singletons would be open.) In fact, there does not seem to be an elegant frame-theoretic correspondent to the DSO condition.

11.3 Derived set semantics of $KD45$

Recall that we are aiming towards a completeness theorem for the doxastic logic $KD45$ with respect to a semantics concerned with the derived set operation. We have seen the axioms of $KD45$ in the opening part of Sec. 11. We now give the semantics. Let $\mathcal{X} = \langle X, \mathcal{O}, i \rangle$ be a *topological model*, a topological space with an interpretation of some fixed background set of atomic sentences.

We interpret the basic modal language in $\mathcal{X}$ using the classical interpretation of the connectives, and most critically,

$$w \models B\phi \quad \text{iff} \quad (\exists U \in \mathcal{O})(w \in U \wedge (\forall x \in U \setminus \{w\})(x \models \phi)).$$

Ax	formal statement	topological correspondent
K	$B(\phi \to \psi) \to (B\phi \to B\psi)$	(none)
D	$B\phi \to \neg B\neg\phi$	dense in itself
4	$B\phi \to BB\phi$	T_d separation property
5	$\neg B\phi \to B\neg B\phi$	all derived sets are open

Figure 6.6. Axiom schemes of $KD45$ again, this time with their topological correspondents.

In words, $B\phi$ is true at w when there is an open set U containing w such that every point in U *except possibly for* w satisfies ϕ. So we have

$$w \models \neg B\neg\phi \text{ iff } (\forall U \in \mathcal{O})(\text{ if } w \in U \text{ then } (\exists x \in U \setminus \{w\})(x \models \phi)).$$

And therefore

$$i(\neg B\neg\phi) \quad = \quad d(i(\phi)).$$

At this point we have a hint as to why the semantics of B was taken with the existential condition rather than the universal one. In Example 6.23 we saw that the derivative operation corresponds to the existential modality $\Diamond$. And since the $KD45$ axioms for B are universal, we take the semantics using the dual of d rather than d itself.

We have the standard semantic definitions: A sentence ϕ is *valid in a topological model* $\mathcal{X}$ iff ϕ is true at every $x \in X$. A sentence ϕ is *valid in a topological space* $\langle X, \mathcal{O}\rangle$ iff ϕ is valid in every topological model based on $\langle X, \mathcal{O}\rangle$.

We first study some correspondence phenomena and then use these to motivate a completeness theorem. The correspondence for D may be found in Shehtman, 1990, and the one for 4 in Esakia, 2001).

THEOREM 6.27 *The correspondences of Fig. 6.6 hold. That is, each scheme is valid on a topological space $\mathcal{X} = \langle X, \mathcal{O}\rangle$ iff $\mathcal{X}$ has the specified topological property.*

Proof We fix a space $\mathcal{X} = \langle X, \mathcal{O}\rangle$ in this proof. The K axioms are easily seen to be valid on $\mathcal{X}$; this amounts to the closure of topologies under intersection.

We turn to the correspondence for the D axioms. If $\{w\}$ is open, consider the model obtained by $i(p) = \{w\}$. In it, $w \models Bp \wedge B\neg p$, counter to D. In the other direction, fix an interpretation i and assume that $w \models B\phi \wedge B\neg\phi$. Then there are opens U and V containing w such that at all points of U besides w, ϕ holds, and all points of V besides w, ϕ fails. Then the open set $U \cap V$ must be $\{w\}$.

Next, suppose that the space is T_d, so every derived set $d(A)$ is closed. Fix i and ϕ, and suppose that $w \models B\phi$. Let $A = i(\neg\phi)$, so that $w \in X \setminus d(A)$. Let U be such that $w \in U \subseteq X \setminus d(A) = i(B\phi)$. U shows that $w \in i(BB\phi)$. Going

the other way, suppose that every 4 axiom holds under every interpretation. Let $A \subseteq X$. We check that $X \setminus d(A)$ is open. Let $i(p) = X \setminus A$, so that $i(Bp) = X \setminus d(A)$. Let $x \in X \setminus d(A)$. So

$$x \in i(Bp) \quad \subseteq \quad i(BBp) \quad = \quad X \setminus (d(X \setminus i(Bp)).$$

Thus there is an open set U containing x with the property that all points in U are in $i(Bp)$, except possibly x. But $i(Bp) = X \setminus d(A)$. And as we know, x itself belongs to $X \setminus d(A)$. So U shows that $X \setminus d(A)$ is indeed open.

For the correspondence result for the 5 axioms, suppose that $d(A)$ is always open in $\mathcal{X}$. Fix i and ϕ, and suppose that $w \models \neg B\phi$. Let $A = i(\neg\phi)$, so that $i(B\phi) = X \setminus d(A)$ and $w \in d(A)$. Then $d(A)$ is open, contains w, and every point of it is outside of $i(B\phi)$. So $w \in i(B\neg B\phi)$. Going the other way, suppose that every 5 axiom holds under every interpretation. Let $A \subseteq X$. We check that $d(A)$ is open. As in the last part, let $i(p) = X \setminus A$. Let $x \in d(A) = i(\neg Bp)$. Then $x \in i(B\neg Bp)$. So there is an open set U containing x such that $U \setminus \{x\} \subseteq i(\neg Bp) = d(A)$. Hence $U \subseteq d(A)$. QED

There are a number of other general facts concerning this semantics which might be of interest. For one, no topological space validates $B\phi \to \phi$. (This is the scheme T which we frowned on before.) But since our main goal is a completeness theorem, we shall not digress.

Recall the notion of a DSO space from Sec. 11.2, and also Proposition 6.24. We see from the correspondence results that every DSO space satisfies all of the $KD45$ axioms under every interpretation. In fact, it is not hard to check that the logic $KD45$ is sound for DSO spaces in the following sense. Let $T \cup \{\phi\}$ be any set of modal sentences (using B as a modality). We say that $T \vdash \phi$ is *provable in* $KD45$ if there is a finite subset $\{\psi_1, \ldots, \psi_n\} \subseteq T$ such that $(\psi_1 \wedge \cdots \wedge \psi_n) \to \phi$ in $KD45$. It is easy to check that $KD45$ is sound for DSO spaces: if $T \vdash \phi$, then $T \models \phi$.

The derived set logic as a fragment of a richer topologic-like system. Much of our chapter has been concerned with the two-sorted ontology of points and sets, and with the resulting bimodal logical systems used in its study. Our work in this section appears to be on a different track. However, it is possible to unify the two. Consider the language $\mathcal{L}$ for points and sets from Sec. 4; this is the language with modalities $\Box$ and K. Instead of K, take as primitive an "everywhere but here" operator $[\neq]$ as in Sec. 3.2, but this time with semantics

$$p, u \models [\neq]\phi \quad \text{iff} \quad \text{for all } q \neq p \text{ in } u, \text{ we have } q, u \models \phi$$

Then the K of topologic is recovered as an abbreviation. Call the resulting language $\mathcal{L}_2$. It is easy to check that the correspondence results of Proposition 6.5 still hold. We can translate the language of this section into $\mathcal{L}_2$ in the manner

of Sec. 4.3 via

$$(B\phi)^* \quad \text{iff} \quad \Diamond[\neq]\phi^*$$

The logic $S4D$ must be modified a bit: the axioms $K\phi \to \Box\phi$ are no longer sound. This reflects the fact that the sentences in $\mathcal{L}_2$ are not always persistent.

Here is a sample of the use of this logic. We saw in Proposition 6.24 the general fact about derived sets $d(d(A)) \subseteq A \cup d(A)$. Here we show that this can be proved in the logic of set spaces together with the "weak 4" scheme $(\phi \wedge [\neq]\phi) \to [\neq][\neq]\phi$. That is, we can prove

$$(\neg p \wedge \Box\langle\neq\rangle\Box\langle\neq\rangle p) \to \Box\langle\neq\rangle p$$

for atomic p. The first conjunct on the left and atomic permanence gives $\Box\neg p$. The second conjunct on the left easily gives $\Box\langle\neq\rangle\langle\neq\rangle p$. Using weak 4 and modal reasoning, we have $\Box\langle\neq\rangle p$.

But we must mention that $\mathcal{L}_2$ has not been studied previously. So all of the natural questions about the logic are open.

11.4 Completeness of $KD45$ for DSO spaces

The final result of our chapter is the following completeness theorem.

THEOREM 6.28 *The logical system $KD45$ is sound and strongly complete for DSO topological models. That is, if $T \models \phi$, then $T \vdash \phi$.*

Proof As the reader expects, the method of proof is to show that a set S of modal sentences which is consistent in $KD45$ has a DSO topological model. So fix such a set S. By the relational completeness noted in Sec. 11.1, we have a model

$$M \quad = \quad \langle W, \to, i\rangle,$$

which is serial, transitive, and Euclidean; and $w^* \in W$ so that $w^* \models S$.

Let N be the set of natural numbers. We build a topological space $\mathcal{M}$ whose set of points is $N \times W$. Here we meld the constructions in Examples 6.22 and 6.23. For the topology, we consider the following family $\mathcal{O}$ of subsets of $N \times W$:

$$U \in \mathcal{O} \qquad \text{iff} \qquad \begin{array}{l}\text{for all } (n,x) \in U, \text{ if } x \to y,\\ \text{then for all but finitely many } m,\ (m,y) \in U.\end{array}$$

It is easy to see that $\mathcal{O}$ is closed under arbitrary unions. So to check that $\mathcal{O}$ really is a topology we only need to consider binary intersections. This amounts to the fact that the intersection of two cofinite sets is again cofinite. As an example, take any $(n,x) \in N \times W$ and consider

$$U_{(n,x)} \quad = \quad \{(n,x)\} \cup \{(m,y) : m \in N \text{ and } x \to y\} \tag{6.8}$$

This is open by the transitivity of $\rightarrow$ in M.

We check that this space $\mathcal{M} = \langle N \times W, \mathcal{O} \rangle$ is a DSO space. No singletons are open, using the fact that the original M is serial. To check that derived sets are open, and also for future use, we again compute derived sets:

$$d(A) \quad = \quad \{(n,x) : n \in N \text{ and there is some } y \text{ such that } x \rightarrow y \text{ and infinitely many } m \text{ such that } (m,y) \in A\}$$

(We leave the verification of this to the reader: see Examples 6.22 and 6.23 for parallel work.) We check that such a set $d(A)$ is open. Suppose that $(n,x) \in d(A)$. Fix some y such that $x \rightarrow y$ and for which there are infinitely many m such that $(m,y) \in A$. Let $U_{(n,x)}$ be as in (6.8). We claim that $U_{(n,x)} \subseteq d(A)$; let $x \rightarrow z$, and consider (p,z). Let V be any open set containing (p,z). The key here is that since $x \rightarrow y$ and $x \rightarrow z$, we have $z \rightarrow y$ by the Euclidean property. Almost all j have the property that $(j,y) \in V$, and infinitely many of them have the property that $(j,y) \in A$. So there is one (indeed infinitely many) j so that $(j,y) \in V \cap A$. Since V is arbitrary, $(p,z) \in d(A)$, as desired.

So far we have converted our relational model M into a DSO space $\mathcal{M}$. We consider it as a model via $i^{\mathcal{M}}(p) = N \times i^{M}(p)$.

LEMMA 6.29 (TRUTH LEMMA) *For all* $(n,x) \in N \times W$, $x \models \phi$ *in* M *iff* $(n,x) \models \phi$ *in* $\mathcal{M}$.

Proof By induction on ϕ. Only the inductive step for B needs an argument. Suppose $x \models B\phi$ in M, so that for all y such that $x \rightarrow y$, $y \models \phi$. By induction hypothesis, we see that for all m, $(m,y) \models \phi$ in $\mathcal{M}$. So each point of $U_{(n,x)}$ from (6.8) above satisfies ϕ in $\mathcal{M}$, except possibly for (n,x). Thus our topological semantics of $B\phi$ tells us that $(n,x) \models B\phi$ in $\mathcal{M}$. Conversely, assume that $(n,x) \models B\phi$ in $\mathcal{M}$. Let U be an open set containing (n,x) with the property that all points in U satisfy ϕ in $\mathcal{M}$, save possibly for (n,x) itself. Let $x \rightarrow y$ in M. There is some m such that $(m,y) \in U$. By induction hypothesis, $y \models \phi$ in M. And since y is arbitrary, we see that in M, $x \models B\phi$. QED

This Truth Lemma completes the proof of Theorem 6.28. We started with a set S and a world $w^* \in W$ such that $w^* \models S$. Then $\mathcal{M}$ is a DSO space, and in it $(0, w^*) \models S$. QED

Incidentally, the same proof shows the following results: $K4$ is complete for spaces in which every derived set is closed (due to Esakia, 2001), and $KD4$ is complete for spaces which are dense in themselves and in which every derived set is closed. For this last result, see Shehtman, 1990 and also Sec. 3 of Ch. 5 of this Handbook. The point is that the construction in the proof of Lemma 6.29 only used the Euclidean property in verifying that the space $\mathcal{M}$ has the property that all derived sets are open. We can drop this point, and

instead verify directly that $\mathcal{M}$ is T_1. The rest of the arguments are the same. Perhaps the progenitor of all of these results is Esakia's completeness theorem for (finite) topological spaces using weak $K4$, the modal logic K supplemented with the scheme $(\phi \wedge B\phi) \to BB\phi)$ (Esakia, 2001). See also Theorem 1.62 of Ch. 5.

12. Other work connected to this chapter

Chapters 5 and 10 contain much material that is relevant to the concerns of this chapter.

We mention a few other papers which also are relevant, but which we did not discuss in the chapter. Davoren, 1999 and Davoren and Gore, 2002 study a propositional bimodal logic consisting of two $S4$ modalities $\Box$ and $[a]$, together with the interaction axiom scheme $\langle a\rangle\Box\phi \to \Box\langle a\rangle\phi$. In the intended semantics, the plain $\Box$ is given the McKinsey-Tarski interpretation of interior, while the labeled $[a]$ is given the standard Kripke semantics using a preorder R_a. The interaction axiom has the flavor of the Cross-Axiom, and here it expresses the property that the R_a relation is lower semi-continuous with respect to the topology.

Pacuit and Parikh, 2005 study a non-spatial application of topologic, thereby showing that the area of application may indeed be wide. They consider a set of agents connected in a communication graph, and such that agent i may receive information from agent j only if there is an edge from i to j. The logic which arises uses a language very similar to *topologic*, and it is shown that for each graph, the logic is decidable, and completely characterizes the graph. An application to the Valerie Plame affair, a notorious political affair from the early years of this century, is also described.

Acknowledgment. The first author's research was supported by a grant from the PSC-CUNY FRAP program. We thank Guram Bezhanishvili, Bernhard Heinemann and Eric Pacuit for comments.

References

Aiello, Marco, van Benthem, Johan, and Bezhanishvili, Guram (2003). Reasoning about space: the modal way. *J. Log. Comput.*, 13(6):889–920.

Barwise, Jon (1988). Three views of common knowledge. In Vardi, M., editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 365–379. Morgan Kaufmann, San Francisco.

Bezhanishvili, Guram, Esakia, Leo, and Gabelaia, David (2005). Some results on modal axiomatization and definability for topological spaces. *Studia Logica*, 81(3):325–355.

Dabrowski, Andrew, Moss, Lawrence S., and Parikh, Rohit (1996). Topological reasoning and the logic of knowledge. *Annals of Pure and Applied Logic*, 78(1–3):73–110. Papers in honor of the Symposium on Logical Foundations of Computer Science, "Logic at St. Petersburg" (St. Petersburg, 1994).

Davoren, J. M. (1999). Topologies, continuity and bisimulations. *Theoretical Informatics and Applications*, 33(4/5):357–381.

Davoren, Jen M. and Gore, Rajeev P. (2002). Bimodal logics for reasoning about continuous dynamics. In *Advances in Modal Logic (Leipzig, 2000)*, volume 3, pages 91–111. World Sci. Publishing, River Edge, NJ.

Esakia, Leo (1981). Diagonal construction, Loeb's formula and Cantor's scattered spaces. In *Logical And Semantical Investigations*, pages 128–143. Academy Press, Tbilisi. In Russian.

Esakia, Leo (2001). Weak transitivity—a restitution. *Logical Investigations*, 8:244–255. In Russian.

Gabelaia, David (2001). Modal definability in topology. Master's thesis, ILLC, University of Amsterdam.

Georgatos, Konstantinos (1993). *Modal Logics for Topological Spaces*. PhD thesis, CUNY Graduate Center.

Georgatos, Konstantinos (1994a). Knowledge theoretic properties of topological spaces. In Masuch, Michael and Laszlo, Polos, editors, *Knowledge Representation and Uncertainty*, Lecture Notes in Comput. Sci. Springer, Berlin.

Georgatos, Konstantinos (1994b). Reasoning about knowledge on computation trees. In *Logics in artificial intelligence (York, 1994)*, volume 838 of *Lecture Notes in Comput. Sci.*, pages 300–315. Springer, Berlin.

Georgatos, Konstantinos (1997). Knowledge on treelike spaces. *Studia Logica*, 59:271–301.

Heinemann, Bernhard (1997). A topological generalization of propositional linear time temporal logic. In *Mathematical Foundations of Computer Science 1997 (Bratislava)*, volume 1295 of *Lecture Notes in Comput. Sci.*, pages 289–297. Springer, Berlin.

Heinemann, Bernhard (1998). Topological modal logics satisfying finite chain conditions. *Notre Dame Journal of Formal Logic*, 39(3):406–421.

Heinemann, Bernhard (1999a). The complexity of certain modal formulas on binary ramified subset trees. *Fundamenta Informaticae*, 39(3):259–272.

Heinemann, Bernhard (1999b). Temporal aspects of the modal logic of subset spaces. *Theoret. Comput. Sci.*, 224(1–2):135–155. Logical foundations of computer science (Yaroslavl, 1997).

Heinemann, Bernhard (2001). Modelling change with the aid of knowledge and time. In *Fundamentals of computation theory (Riga, 2001)*, volume 2138 of *Lecture Notes in Comput. Sci.*, pages 150–161. Springer, Berlin.

Heinemann, Bernhard (2006). Regarding overlaps in "topologic". In Hodkinson, I. and Venema, Y., editors, *Advances in Modal Logic, AiML 2006, Noosa, Queensland, Australia*, volume 6. King's College Publications, London.

Kozen, Dexter and Parikh, Rohit (1981). An elementary proof of the completeness of PDL. *Theoretical Computer Science*, pages 113–118.

Krommes, G. (2003). A new proof of decidability for the modal logic of subset spaces. In *Eighth ESSLLI Student Session*, pages 137–148.

Kudinov, Andrey (2006). Topological modal logics with difference modality. In Hodkinson, I. and Venema, Y., editors, *Advances in Modal Logic, AiML 2006, Noosa, Queensland, Australia*, volume 6. King's College Publications, London.

McKinsey, J. C. C. (1941). A solution of the decision problem for the lewis systems S2 and S4, with an application to topology. *J. Symbolic Logic*, 6: 117–134.

McKinsey, J. C. C. and A. Tarski (1944). The algebra of topology. *Annals of Mathematics*, 45:141–191.

Moss, Lawrence S. and Parikh, Rohit (1992). Topological reasoning and the logic of knowledge. In Moses, Y., editor, *Theoretical Aspects of Reasoning About Knowledge*, pages 95–105. Morgan Kaufmann.

Pacuit, Eric and Parikh, Rohit (2005). The logic of communication graphs. In Leite, J., Omicini, A., Torroni, P., and Yolum, P., editors, *Declarative Agent Languages and Technologies II: Second International Workshop, DALT 2004*, volume 3476 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin.

Shehtman, V. (1990). Derived sets in euclidean spaces and modal logic. Technical Report X-90-05, University of Amsterdam.

Steinsvold, Chris (2006). *Topological Models of Belief Logics*. PhD thesis, CUNY Graduate Center.

van Benthem, Johan, Bezhanishvili, Guram, ten Cate, Balder, and Sarenac, Darko (2007). Multimodal logics of products of topologies. *Studia Logica*. to appear.

van Benthem, Johan and Sarenac, Darko (2004). The geometry of knowledge. Technical Report PP-2004-20, ILLC.

Vickers, Steven (1989). *Topology via Logic*. Cambridge University Press, Cambridge.

Weiss, M. A. (1999). *Completeness of Certain Bimodal Logics*. PhD thesis, CUNY Graduate Center.

Weiss, M. A. and Parikh, R. (2002). Completeness of certain bimodal logics for subset spaces. *Studia Logica*, 71(1):1–30.

Chapter 7

LOGICAL THEORIES FOR FRAGMENTS OF ELEMENTARY GEOMETRY

Philippe Balbiani
Institut de Recherche en Informatique de Toulouse

Valentin Goranko
University of the Witwatersrand

Ruaan Kellerman
University of Johannesburg

Dimiter Vakarelov
University of Sofia

Second Reader

Yde Venema
University of Amsterdam

1. Introduction and historical overview

In ancient Babylon and Egypt geometry was just a set of empirical observations and practical skills and methods for measuring land and designing irrigation systems, although it already had a degree of sophistication (e.g., Pythagoras' theorem was already known and the triangle with sides 3-4-5 was used in practice for producing right angles). In ancient Greek times it evolved into a *liberal art*.

We dare claim that Geometry only became a science with Euclid's epic work "Elements" written over 2300 years ago, which was not only the first truly

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 343–428.
© 2007 *Springer.*

scientific treatment of geometry, but also the first systematic application of the *axiomatic method* in mathematics. Geometry remained a central subject in mathematics throughout the centuries, but when in the first half of the 17th century Descartes introduced coordinate systems, and with them the *analytic method* in geometry, it gradually began to lose its prime position in mathematics and became part of algebra and calculus, and later—of topology. The modern view, going back to the famous Klein's *Erlangen program*, defined geometry as a study *not of figures, but of transformations,* and classified the different geometric structures and their theories in terms of the groups of transformations which preserve them. This view placed it firmly on algebraic foundations to an extent that some mathematicians consider it as an "applied group theory".

On the other hand, the discovery of non-Euclidean geometries by Bolyai, Lobachevsky and Gauss in the early 19th century (see e.g. Coxeter, 1969, Eves, 1972, Meserve, 1983), which showed *inter alia* the independence of Euclid's "Fifth Postulate" from the other axioms of Euclidean geometry, was an impressive demonstration of the strength and importance of the formal logical approach in mathematics, which also reinforced the importance of geometry to the foundations of mathematics. Euclid's Fifth Postulate claims that, given a line and a point not incident with it in a plane, there exists a unique line in that plane passing through the given point and parallel to the given line. Depending on the acceptance or otherwise of that postulate, several natural lines of development of geometry evolve:

- *affine geometry*, first studied by Euler, which adopts the Fifth Postulate. Thus incidence, parallelism, collinearity, and betweenness, as well as transformations that preserve these relations, play a central role in affine geometry. Such *affine transformations* can be taken, in the spirit of the Erlangen program, as defining the very notion of "affine". Affine geometry does not deal with angles, distances, or any other related metric concepts (not even with orthogonality), as these are not invariant under affine transformations. Thus the models of affine geometry, viz. affine spaces, are more general than Euclidean spaces, but have poorer structure.

- *hyperbolic geometry*, introduced by Lobachevsky and Bolyai, and *elliptic* geometry, which adopt the negation of the Fifth Postulate. In hyperbolic geometries, given a line and a point not on that line, there exist infinitely many lines through that point which lie parallel with the original given line (see e.g. Coxeter, 1969, Szczerba and Tarski, 1965, Szczerba and Tarski, 1979 for more details); in elliptic geometry (see Coxeter, 1969, Behnke et al., 1974) there are no parallel lines at all. These will not be discussed in this chapter, but see Hilbert, 1950 for comparison and relationships.

- *absolute geometry*, introduced by J. Bolyai, based only on the first four postulates of Euclid, but independent of the Fifth Postulate. In some extension of absolute geometry the notion of parallelism can be completely rejected, as in *projective geometry*, where every two lines in a projective plane intersect. In others, alternatives of the Fifth Postulate can be adopted, as in the elliptic and hyperbolic geometries. Thus, absolute geometry is a full-fledged system of geometry, involving distances, angles, etc., but based on a weaker axiomatic basis than Euclidean geometry. It is not a subsystem of affine geometry, though the intersection of the two is still rich enough to develop a meaningful and interesting theory of ordered affine structures (see Coppel, 1998, Coxeter, 1969, Ch. 15, Lenz, 1992, Szczerba, 1972).

A fundamental affine relation, i.e., one invariant under affine transformations, is the relation of *betweenness* on triples of points, which extends basic affine structures by introducing *ordering* between points on a line, but not distances. Much of the theory of betweenness is independent of Euclid's Fifth Postulate, and thus lies in the intersection of absolute and affine geometry (see Coppel, 1998, Coxeter, 1969).

Perhaps the simplest important non-affine relation is that of *orthogonality*. Further adding distances and angles, along with the axioms of the field of reals, extends affine geometry to the classical, Euclidean geometry of the real plane and space.

At the beginning of the 20th century, Hilbert, the most influential proponent of the axiomatic method in mathematics, illustrated the power of that method by re-casting Euclid's work into a precise and rigorous modern treatment which eventually put geometry on sound axiomatic foundations (see Hilbert, 1950). It was preceded by axiomatic investigations of the foundations of geometry at the end of the 19th century by Peano, as well as Pieri, 1908, Veblen, 1904, Veblen, 1914 and Pasch, 1882, who analyzed various axiomatic systems and the mutual relationships between the primitive notions of Euclidean geometry. However, the axiomatic method in geometry only reached its logical maturity with the seminal work of Tarski and his students and followers Szczerba, Szmielew, Schwabhäuser, Scott, Monk, Givant and others (see Schwabhäuser et al., 1983 for comprehensive details) in the 1920-70's. Tarski developed systematically the logical foundations of *elementary geometry*, which is "that part of Euclidean geometry that can be formulated and established without the help of any set-theoretical devices" (see Tarski, 1959). Essentially, that means the first-order theory of Euclidean geometry, developed over a suitably expressive first-order language (see below). In particular:

- Tarski, 1951, Tarski, 1967 demonstrated how the elementary geometry of the real plane can be formally interpreted into the elementary (i.e.

first-order) theory of real-closed fields. Furthermore, Tarski showed the completeness and decidability of the theory of real-closed fields by means of quantifier elimination, and consequently obtained a decision procedure for the elementary Euclidean geometry. He then extended these results to Euclidean planes over arbitrary real-closed fields.

- Tarski, 1959 showed that the whole of elementary geometry can be developed axiomatically using just two geometric relations, viz. *betweenness* and *equidistance* (used as the only primitives also by Veblen, 1904). He thus obtained an explicit axiomatization of the first-order theory of the Euclidean geometry in terms of these primitives and showed that it is complete and decidable, though not finitely axiomatizable.

- In a similar fashion, Szmielew, 1959 studied the first-order theory of the metric hyperbolic geometry, obtained by negating Euclid's axiom in Tarski's first-order axiomatization of the Euclidean geometry.

- Szczerba and Tarski, 1965, Szczerba and Tarski, 1979 studied and characterized the first-order theories of the fragments of the Euclidean, hyperbolic and absolute geometries based on betweenness alone, for which they established explicit axiomatizations.

- Beth and Tarski, 1956, Tarski, 1956 studied the problem of which geometric relations are sufficient to be adopted as primitive notions in terms of which the whole of Euclidean geometry can be developed.

- Szmielew, 1983 developed the theory of point-based collinearity structures and showed how to build up the Euclidean geometry from that theory, while Schwabhäuser and Szczerba, 1975 studied line-based structures for elementary geometry.

While the post-Tarski period in the logical foundations of geometry is less active and spectacular, still there are several research lines which deserve discussion. Besides the notable works of Tarski's students mentioned above, they include:

- the study of classical and constructive axiomatizations of fragments of projective, affine, absolute, Euclidean, elliptic, hyperbolic, etc. geometries, with emphasis on simplicity and minimality, in Pambuccian, 1989, von Plato, 1995, Lombard and Vesley, 1998, Pambuccian, 2001a, Pambuccian, 2001b, Pambuccian, 2006, etc. For a general discussion of the axiomatics of affine and projective geometry, see Bennett, 1995.

- the investigation of primitive relations sufficient for the elementary affine, projective, absolute, etc. geometries; the expressiveness of such relations;

and axiomatizations in terms of such relations, in Scott, 1956, Pambuccian, 1995, Pambuccian, 2003, Pambuccian, 2004, etc.

- the development of practical methods and algorithms for theorem proving in algebra and geometry: quantifier elimination based methods, such as Seidenberg's implementation of Tarski's method, Seidenberg, 1954, the method of cylindrical algebraic decompositions (see e.g. Caviness and Johnson, 1998, Buchberger et al., 1988), the more recent and efficient Heintz et al., 1990, Renegar, 1992, Basu et al., 1996, Basu, 1999, the Gröbner basis method (Buchberger, 1985), the characteristic set method (Chou and Gao, 1990) and others. For more details and references see Sec. 9.2.

Most of the studies and results mentioned above apply to geometric structures of which the logical languages are rich enough to express properties of ordering and metric. However, there are various weaker, yet natural and important, geometric structures such as *parallelism*, *orthogonality*, *incidence* and *collinearity structures*, which involve points and lines in a real or abstract geometric space. The elementary theories of these latter structures are considerably less studied, mainly from the perspective of discrete and combinatorial geometry. We will discuss these structures and their theories in some detail here, as they play an important role in various models of qualitative spatial reasoning.

While affine and absolute spaces are too general to allow the development of a full-fledged elementary geometry in them, they are still amenable to algebraic treatment by means of *coordinatization*, which enables the study of affine and projective spaces by studying algebraic structures called *ternary rings* (see e.g. Blumenthal, 1961, Heyting, 1963, Szmielew, 1983, Hughes and Piper, 1973, Mihalek, 1972). Since the coordinatization is a first-order interpretation, it is instrumental for the algebraic investigation and characterization of the logical theories of affine spaces, and can be used to establish various logical properties, such as independence results, representation theorems, (lack of) finite model property, decidability and complexity results of these theories.

In this chapter we survey and discuss from a *logical perspective* structures and theories of parallelism, orthogonality, incidence and order, gradually building the full *elementary geometry* of Euclidean spaces, in Tarski's sense. Besides traditional geometric properties and constructions, we discuss various logical issues such as: *definability of relations and properties*, *expressiveness of concepts*, *axiomatic theories and their models*, *representation results and completeness*, *finite model property*, *decidability*, *categoricity* and other model-theoretic properties.

The chapter consists of two parts. In the first part we discuss classical, first-order theories of geometric structures, starting with very weak structures of parallelism (Sec. 3), orthogonality (Sec. 4), incidence (Sec. 5) and collinearity,

for which we show how to develop some geometric concepts, such as independence, basis, planarity and dimension. In Sec. 6 and Sec. 7 we outline coordinatization of projective and affine planes as a general method of interpreting them into algebraic structures called planar ternary rings, and discuss the relationship between geometric and algebraic properties, and generally between the logical theories of planes and the associated coordinate rings. We also discuss collineations and general affine transformations, and the associated (invariant) affine concepts and properties. In Sec. 8 we then add betweenness and order in affine planes, discuss definability in these planes, and the relationship of these planes with ordered coordinate rings, as well as the results from Szczerba and Tarski, 1965, Szczerba and Tarski, 1979 on axiomatic theories of betweenness. Eventually we consider some rich languages, i.e. languages containing primitive notions in terms of which the whole of elementary geometry can be developed, and present Tarski's axiomatization of the Euclidean geometry in terms of betweenness and equidistance. The first part of the chapter, dealing with elementary theories of geometry, ends with a brief discussion in Sec. 9.2 of the development of decision methods for elementary geometry since Tarski's seminal decidability results, and automated reasoning for elementary geometry.

The second part of the chapter is devoted to *modal logics* arising from classical, mainly two-dimensional geometrical structures. After a short general discussion of spatial modal logics in Sec. 10, we consider modal logics of several sorts: *point-based* (Sec. 11), *line-based* (Sec. 12), and *point-line based logics* (Sec. 14) with incidence relations between the sorts, defining affine or projective incidence structures. In Sec. 13 we show how two-sorted relational structures based on points and lines can be replaced by one-sorted relational structures containing the same geometrical information, and how modal logic can be developed on such structures. In Sec. 14 we discuss point-line spatial logics and show how modal languages can be interpreted on two-sorted relational structures.

2. Preliminaries

2.1 Some terminology and notation

The following notions will be introduced more than once in this chapter. Here we only fix the notation and terminology used further (unless otherwise specified) for the convenience of the reader.

We deal with two basic geometric objects: *points* and *lines*.

Points. Points are usually considered primitive concepts, but as we shall see, they can also be defined in terms of co-punctual lines. Specific points will be denoted as A, B, C etc. and typical point variables will be X, Y, Z, etc. Basic

relations on points are *collinearity*, denoted as $\mathbf{Col}(XYZ)$, meaning that the points X, Y and Z lie on a common line (sometimes generalized to n points); *betweenness*, denoted as $\mathbf{B}(XYZ)$, meaning that Y lies on the line segment joining X and Z (with possibly Y coinciding with X or Z); and *equidistance*, denoted $XY \equiv ZU$ and meaning that the line segment formed by X and Y has the same length as the line segment formed by Z and U. Using any of these, one can define the *triangle* relation, which holds when three points are non-collinear.

Lines. Lines can be introduced as primitive concepts, or defined in terms of pairs of different points, or as equivalence classes of points in collinearity structures. Specific lines will be denoted as a, b, c, etc. Typical line variables will be x, y, z etc. Basic relations on lines are:

Incidence, denoted as $x\,\mathbf{Inc}\,y$, meaning that the lines x and y share a common point, and may even coincide. Incidence may be generalized to *co-punctuality* (*concurrence*), denoted $\mathbf{Cop}(x_1 \ldots x_n)$, meaning that the lines $x_1, \ldots, x_n$ have exactly one point in common.

Intersection of two lines, denoted $x\ \mathbf{Int}\ y$ and meaning that x and y are incident *but different*.

Co-planarity of lines, denoted $\mathbf{Pl}(xy)$ for two lines and generalized to $\mathbf{Pl}(x_1 \ldots x_n)$ for n lines, meaning that the lines lie in the same plane.

Strict (irreflexive) parallelism, denoted $x \parallel y$, meaning that the lines x and y are parallel and different, and *weak (reflexive) parallelism*, denoted $x \mathrel{\shortparallel} y$, meaning that x and y are parallel or coincide.

Orthogonality, denoted $x \perp y$, meaning that the lines x and y are orthogonal, but not necessarily intersecting, and *perpendicularity*, denoted $x \mathrel{\dot{\perp}} y$, meaning that the lines x and y are orthogonal and coplanar (and hence intersecting).

Skewness, denoted $x \bowtie y$, meaning that x and y are not co-planar.

Lines can also be defined as sets of points collinear with a pair of points: given two distinct points P and Q, the *line determined by P and Q*, denoted $\mathbf{l}(P, Q)$, is defined as the set of all points X such that $\mathbf{Col}(PQX)$ holds.

Given a point X and a line y, the claim that X is *incident* with y will be denoted as $X\mathbf{I}y$ or simply as $X \in y$, while, assuming Euclid's parallel postulate, the unique line parallel with y and containing X will be denoted $\mathbf{p}(X, y)$. The unique line incident with two distinct points X and Y will be denoted as $\mathbf{l}(X, Y)$ or simply as XY, while the line segment between X and Y will be denoted $|XY|$ and the length of that segment as $\|XY\|$. Given intersecting lines x and y, their point of intersection will be denoted $\mathbf{P}(x, y)$.

For every integer $n \geq 1$, $\mathrm{Diff}_n(X_1 \ldots X_n)$ will be the formula stating that $X_1, \ldots, X_n$ are distinct, i.e.

$$\mathrm{Diff}_n(X_1 \ldots X_n) \ := \bigwedge_{\substack{i \neq j \\ 1 \leq i,j \leq n}} X_i \neq X_j,$$

and likewise for $\mathrm{Diff}_n(x_1 \ldots x_n)$.

2.2 Algebraic background

The terminology on algebraic structures varies considerably in the literature, so we fix ours here. The reader is referred to any standard text in abstract algebra, or to Szmielew, 1983 for more details.

Consider the structure $(G; 0, +)$ where G is a non-empty set, $+$ is a binary operation on G and 0 is some distinguished element in G. Then $(G; 0, +)$ is called an (*additive*) *loop* (*with zero* 0) if

1. $a + 0 = a = 0 + a$ for all $a \in G$;
2. $a + b = c$ uniquely determines any of a, b, c from the other two.

Likewise we refer to $(G; 1, \cdot)$ as a *multiplicative loop with unit* 1.

If the operation $+$ is associative as well, then $(G; 0, +)$ is called a *group*. Note that the second property above guarantees the existence of additive inverses in any group. A group will be called *abelian* when the operation defined in it is commutative.

A structure $(F; 0, +, \cdot)$ is called a *ring* if $(F; 0, +)$ is an abelian group and the multiplication $\cdot$ is both associative and distributive over $+$.

A structure $(G; 0, 1, \cdot)$ (where 0 and 1 are distinct distinguished elements from G) is called a *multiplicative loop with zero* if

1. $(G \backslash \{0\}, 1, \cdot)$ is a multiplicative loop with unit 1;
2. $a \cdot 0 = 0 = 0 \cdot a$ for all $a \in G$.

Here 0 is the zero of $(G; 0, 1, \cdot)$ and 1 is its unit. Again $(G; 0, 1, \cdot)$ will be called a *multiplicative group with zero* when $\cdot$ is associative.

$(F; 0, 1, +, \cdot)$ is called a *double loop* when

1. $(F; 0, +)$ is an additive loop;
2. $(F; 0, 1, \cdot)$ is a multiplicative loop with zero.

A double loop $(F; 0, 1, +, \cdot)$ is called a *left division ring* (respectively, *right division ring*) when $(F; 0, +)$ is an abelian group and $\cdot$ is left-distributive (respectively, right-distributive) over $+$. A *division ring* is a double loop that is

both a left and right division ring. A division ring $(F; 0, 1, +, \cdot)$ with associative multiplication $\cdot$ is called a *skew field*, and if $\cdot$ is also commutative then $(F; 0, 1, +, \cdot)$ becomes a *field*. Note that there is some variation in the literature regarding these terms; for example, division rings are called in Szmielew, 1983 quasi-fields. By a classical result of Wedderburn, every finite skew field is a field.

A structure $(G; 0, +, \leq)$ will be called an *ordered loop* if $(G; 0, +)$ is a loop and $\leq$ is a linear ordering on the set G such that

1. $a \leq b \Rightarrow c + a \leq c + b$ (left additive monotony)

2. $a \leq b \Rightarrow a + c \leq b + c$ (right additive monotony)

for all $a, b, c \in G$. If $(G; 0, +)$ is a group then $(G; 0, +, \leq)$ will be called an *ordered group*, etc.

A structure $(F; 0, 1, +, \cdot, \leq)$ will be called an *ordered double loop* if $(F; 0, 1, +, \cdot)$ is a double loop and $\leq$ is a linear ordering on the set F such that both left and right additive monotony holds, and

1. $a \leq b \Rightarrow c \cdot a \leq c \cdot b$ (left multiplicative monotony)

2. $a \leq b \Rightarrow a \cdot c \leq b \cdot c$ (right multiplicative monotony)

for all $a, b, c \in G$ with $c \geq 0$. Likewise if $(F; 0, 1, +, \cdot)$ is instead, say, a left division ring, then $(F; 0, 1, +, \cdot, \leq)$ will be called an *ordered left division ring*, etc. It is easy to see that if $(F; 0, 1, +, \cdot)$ is an ordered double loop then $0 < 1$ and hence F has infinite cardinality since for every $x \in F$, $x < x + 1$.

An ordered structure $(F; 0, 1, +, \cdot, \leq)$ is *Euclidean*, if for every $a \in F$ with $a \geq 0$ there exists $b \in F$ such that $a = b^2$; *real closed*, if it is Euclidean and every polynomial of odd degree over F has a zero in F.

2.3 Logical background

In the treatment of some logical issues, we assume that the reader has background on the basic model theory of first-order logic, suitable references on which include Doets, 1996, Enderton, 1972 and the very comprehensive and more advanced Hodges, 1993. Here we only mention a few more specific concepts and results used in the chapter.

Theories. A (first-order) theory is any set of first-order sentences. A theory T is *complete* if every two models of the theory are elementarily equivalent, i.e., satisfy the same first-order sentences. A typical example of a complete theory is the set $\mathrm{TH}(\mathfrak{A})$ of all first-order sentences satisfied in a given structure $\mathfrak{A}$. A

theory T is *ω-categorical* (or, countably categorical) if all countable models of T are isomorphic; T is *decidable* if there is an algorithm which can determine if a given sentence is a logical consequence of T. By the Łoś-Vaught Test (see e.g. Doets, 1996) every ω-categorical theory is complete and decidable.

Padoa's method. Let $\mathcal{L}$ be a first-order language over some signature S, let $\mathbf{s}$ be a symbol not in S and T a theory over the signature $S \cup \{\mathbf{s}\}$. If $\mathfrak{A}$ and $\mathfrak{B}$ are models of T with $\mathfrak{A}|S = \mathfrak{B}|S$ but $\mathbf{s}^{\mathfrak{A}} \neq \mathbf{s}^{\mathfrak{B}}$ then $\mathbf{s}$ cannot be defined by T in $\mathcal{L}$. More generally, if $\mathfrak{A}$ is a model of the theory T and if there exists an automorphism of $\mathfrak{A}|S$ that fails to preserve the symbol $\mathbf{s}$, then $\mathbf{s}$ cannot be defined by T in $\mathcal{L}$. For example, consider the structure $(\mathbb{Z}; +)$. The constant $\mathbf{0}$ is explicitly definable using the formula $\varphi_{\mathbf{0}}(x) := \forall y(x + y = y)$. From this we can then explicitly define subtraction using the formula $\chi_{-}(x, y, z) := \exists u \exists v(\varphi_{\mathbf{0}}(u) \wedge y + v = u \wedge x + v = z)$. To show that multiplication $\cdot$ is not definable in $(\mathbb{Z}; +)$, simply note that the automorphism h of $(\mathbb{Z}; +)$ given by $h(x) = -x$ does not preserve multiplication, since in general $-(x \cdot y) \neq (-x) \cdot (-y)$.

Interpretations. Let S be any signature with $S_A, S_B \subseteq S$, and consider the structures $\mathfrak{A} = (A; S_A)$ and $\mathfrak{B} = (B; S_B)$. An *n-dimensional interpretation* of $\mathfrak{B}$ in $\mathfrak{A}$ consists of the following (the vectors $\bar{x}$ will refer to n-tuples of variables):

1. A formula $\varphi(x_1, \ldots, x_n)$ over the signature S_A which defines some relation $D_B \subseteq A^n$ representing the domain B interpreted in A;
2. A bijective ("decoding") function $f : D_B \to B$;
3. (i) for every constant symbol $\mathbf{c} \in S_B$, a formula $\varphi_{\mathbf{c}}(\bar{x})$ over S_A that defines some element $c_B \in D_B$ such that $f(c_B) = \mathbf{c}^{\mathfrak{B}}$;
   (ii) for every m-ary relation symbol $\mathbf{r} \in S_B$, a formula $\varphi_{\mathbf{r}}(\bar{x}_1, \ldots, \bar{x}_m)$ over S_A that defines some relation $r_B \subseteq D_B^m$ such that $f[r_B] = \mathbf{r}^{\mathfrak{B}}$ (likewise for the equality symbol);
   (iii) for every m-ary function symbol $\mathbf{g} \in S_B$, a formula $\varphi_{\mathbf{g}}(\bar{x}_1, \ldots, \bar{x}_m, \bar{x}_{m+1})$ over S_A that defines some function $g_B : D_B^m \to D_B$ such that $f(g_B(\bar{a}_1, \ldots, \bar{a}_m)) = \mathbf{g}^{\mathfrak{B}}(f(\bar{a}_1), \ldots, f(\bar{a}_m))$.

The above definition can be extended to non-injective interpretations and interpretations between many-sorted structures, see e.g. Hodges, 1993. A classical example is the 2-dimensional interpretation of the rationals $\mathfrak{Q} = (\mathbb{Q}; +, \cdot)$ in the integers $\mathfrak{Z} = (\mathbb{Z}; +, \cdot)$, as ordinary fractions. Interpretations will be used in Sec. 7, where affine planes will be interpreted in algebraic structures called ternary rings.

3. Structures and theories of parallelism

We begin our study with very weak and simple structures which consist of a set of lines subject only to the relation of parallelism (besides equality). We

will provide a definitive axiomatic description of such structures which can be extracted from the real Euclidean space of any dimension. In particular, it will turn out that the relation of line parallelism is too weak to distinguish dimensions greater than $n = 1$.

By a *line parallelism frame*, or simply a *parallelism frame*, we mean any structure of the form $\langle \mathbf{Li}, \| \rangle$, where $\|$ is a binary relation called *parallelism* over a non-empty set $\mathbf{Li}$ of which the elements are called *lines*. When the relation $\|$ holds for two lines x and y, we will use phrases such as x *is parallel to* y, etc.

A *pre-model of parallelism* is a parallelism frame $\langle \mathbf{Li}, \| \rangle$ satisfying the following conditions:

$\mathsf{Sym}_{\|}$: $\forall x \forall y (x \parallel y \rightarrow y \parallel x)$ (symmetry)

$\mathsf{PTran}_{\|}$: $\forall x \forall y \forall z (x \parallel y \wedge y \parallel z \rightarrow x = z \vee x \parallel z)$ (pseudo-transitivity)

A pre-model of parallelism in which the parallelism relation is reflexive (respectively, irreflexive) will be called a *model of weak parallelism*, (respectively, a *model of strict parallelism*). Thus, models of weak and strict parallelism must satisfy respectively the axioms:

$\mathsf{Ref}_{\|}$: $\forall x \, (x \parallel x)$, and $\mathsf{Irr}_{\|}$: $\neg \exists x \, (x \parallel x)$.

Hereafter, unless otherwise specified, by parallelism we will mean strict parallelism, and weak parallelism will be denoted by the symbol $\shortparallel$. Clearly, these are definable in terms of one another:

$$x \shortparallel y \;\Leftrightarrow\; x \parallel y \vee x = y, \qquad x \parallel y \;\Leftrightarrow\; x \shortparallel y \wedge x \neq y.$$

Thus models of weak parallelism are simply equivalence relations, while models of strict parallelism are isomorphic to disjoint unions of relational structures of the form $\langle \mathbf{W}, \neq \rangle$, where $\neq$ is the difference relation over some non-empty set $\mathbf{W}$. Given a model of strict parallelism $\langle \mathbf{Li}, \| \rangle$ and any line $x \in \mathbf{Li}$, the set of lines $\{x\} \cup \{y \in \mathbf{Li} : y \parallel x\}$ will be called the *parallel class (containing x)*. The property that a model of strict parallelism contains infinitely many parallel classes can be modelled using the scheme $\mathsf{Par}_{\|}$ consisting of the axioms

$$\exists x_1 \ldots \exists x_k \left(\mathrm{Diff}_k(x_1 \ldots x_k) \wedge \bigwedge_{i \neq j} x_i \nparallel x_j \right)$$

for every natural $k \geq 1$.

In models of strict parallelism, it is possible for a line to be parallel with no other line. A model of parallelism will be called *k-serial* if it satisfies the property

$$\forall x \exists y_1 \ldots \exists y_k \left(\mathrm{Diff}_k(y_1 \ldots y_k) \wedge \bigwedge_{i=1}^{k} y_i \parallel x \right).$$

A model that is k-serial for every natural $k \geq 1$ will be called *infinitely serial*, and the scheme specifying that a model is infinitely serial, consisting of all the above axioms for $k \geq 1$, will be denoted as $\mathsf{Ser}_{\parallel}$.

A model of strict parallelism is *real* if it consists of (not necessarily *all*) lines in the real plane, with the usual relation of strict parallelism.

Given a line u, let $\mathbf{u}$ denote the parallel class of u. Now, with every such parallel class $\mathbf{u}$ we associate a real number $m_{\mathbf{u}}$ meant to represent the slope of the lines in $\mathbf{u}$ in some arbitrarily fixed orthogonal coordinate system in the real plane, so that the mapping m is to be injective. Then, each line v in the class $\mathbf{u}$ can be mapped to a unique real number b_v, and the line v is identified with the line in the real plane having equation $y = m_{\mathbf{u}}x + b_v$. Thus, we have the following elementary characterization of line parallelism in $\mathbb{R}^n$.

PROPOSITION 7.1 *Every model of strict parallelism of cardinality not greater than the continuum is isomorphic to a real model.*

By taking the mappings m and b to be surjective, after setting aside a parallel class for the vertical lines, we obtain:

PROPOSITION 7.2 *Every model of strict parallelism in which there are continuum many parallel classes, each of them with the cardinality of the continuum, is isomorphic to the model of strict parallelism consisting of all lines in* $\mathbb{R}^2$.

COROLLARY 7.3 *For every natural* $n \geq 2$, *the model of strict parallelism consisting of all lines in* $\mathbb{R}^n$ *is isomorphic to the model of strict parallelism consisting of all lines in* $\mathbb{R}^2$.

PROPOSITION 7.4 *The theory of the class of models of strict parallelism which satisfy the schemes* $\mathsf{Par}_{\parallel}$ *and* $\mathsf{Ser}_{\parallel}$ *is* ω*-categorical, and hence complete and decidable.*

Indeed, let $\mathfrak{A}$ and $\mathfrak{B}$ be two countable models of strict parallelism satisfying the schemes $\mathsf{Par}_{\parallel}$ and $\mathsf{Ser}_{\parallel}$. Then $\mathfrak{A}$ contains countably many parallel classes, each of them of countable cardinality, and likewise for $\mathfrak{B}$. Let φ be any bijection between the parallel classes of $\mathfrak{A}$ and the parallel classes of $\mathfrak{B}$, and for every parallel class $\mathbf{x}$ in $\mathfrak{A}$, let $\psi_{\mathbf{x}}$ be a bijection between the lines in $\mathbf{x}$ and the lines in $\varphi(\mathbf{x})$. Then the line u in $\mathfrak{A}$ lying in the parallel class $\mathbf{x}$ is mapped to the line $\psi_{\mathbf{x}}(u)$ in $\mathfrak{B}$, and this establishes an isomorphism between $\mathfrak{A}$ and $\mathfrak{B}$.

The completeness and decidability now follow by the Łoś-Vaught Test.

Since the theory of strict parallelism is reducible to the first-order theory of equality, from Stockmeyer, 1977 it follows that the theory of strict parallelism is PSPACE-complete.

4. Structures and theories of orthogonality

4.1 Orthogonality frames and dimension

By a *line orthogonality frame*, or simply *orthogonality frame*, we will mean any structure of the form $\langle \mathbf{Li}, \perp \rangle$, where $\mathbf{Li}$ is a set, the elements of which will be called *lines*, and $\perp$ is a binary relation on $\mathbf{Li}$, called the *orthogonality* relation. Lines x and y satisfying $x \perp y$ will be called *orthogonal*. If x and y are both orthogonal as well as incident, then they will be called *perpendicular*, denoted $x \dot{\perp} y$. For $n \geq 1$, dimension can be defined in an orthogonality frame using the conjunction of the sentences $\mathsf{dim}_{\perp}^{(n)}$ and $\mathsf{Dim}_{\perp}^{(n)}$, given as

$\mathsf{dim}_{\perp}^{(n)}$: $\exists x_1 \ldots \exists x_n \left(\bigwedge_{i \neq j} x_i \perp x_j \right)$;

$\mathsf{Dim}_{\perp}^{(n)}$: $\neg \exists x_1 \ldots \exists x_n \exists x_{n+1} \left(\bigwedge_{i \neq j} x_i \perp x_j \right)$.

Clearly $\mathsf{Dim}_{\perp}^{(n)} = \neg \mathsf{dim}_{\perp}^{(n+1)}$. An orthogonality frame satisfies the property of *n-dimensionality* (for $n \geq 1$) if both sentences $\mathsf{dim}_{\perp}^{(n)}$ and $\mathsf{Dim}_{\perp}^{(n)}$ hold in that frame. A frame that satisfies 2-dimensionality will also be called *planar*.

When dealing with orthogonality frames, the expression $x_1 \parallel x_2$ will be an abbreviation for the formula

$$\forall y \, (y \perp x_1 \leftrightarrow y \perp x_2) \, . \tag{7.1}$$

In the context of orthogonality frames, by parallelism we will mean weak parallelism, and will say that lines x_1 and x_2 are *parallel* when $x_1 \parallel x_2$ in the sense of (7.1). Clearly, the binary relation defined by $\parallel$ is an equivalence relation.

From the class of all orthogonality frames, we single out those which satisfy the additional axiom

$\mathsf{Pen}_{\perp}$: $\forall x_1 \forall x_2 \, (x_1 \neq x_2 \rightarrow x_1 \nparallel x_2)$.

Such orthogonality frames will be called *pencils*, by analogy with a pencil being a collection of co-punctual lines. However, our pencils of orthogonality are not pencils in the strict sense. The axiom $\mathsf{Pen}_{\perp}$ simply states that there may be no parallel lines, and this mimics pencil structure. But the axioms do not exclude models where the lines are not all co-punctual. In fact, incidence is not even definable from orthogonality, so that it is futile to try and axiomatize co-punctuality of lines in orthogonality frames. For example, let R_M^n be the set of all lines in the Euclidean space $\mathbb{R}^n$, and define $\mathfrak{R}_M^n := (R_M^n; \perp)$, where $\perp$ is the Euclidean orthogonality relation. By using the method of Padoa (i.e. finding an automorphism of $\mathfrak{R}_M^n$ which fails to preserve incidence) we can obtain the following.

PROPOSITION 7.5 *For $n \geq 3$, the relation of line incidence is not definable in $\mathfrak{R}_M^n$.*

Orthogonality pencils can be obtained from orthogonality frames by factoring over parallel classes: if $\mathfrak{A}$ is any orthogonality frame, then $\mathfrak{A}/_{\shortparallel}$, the quotient structure of $\mathfrak{A}$ induced by the parallelism relation defined by (7.1), is an orthogonality pencil; parallelism reduces to equality in orthogonality pencils.

We will call an n-dimensional orthogonality frame *real* if it can be isomorphically embedded in $\mathbb{R}^n$ with the Euclidean orthogonality relation, where two lines are orthogonal when the dot product of their direction vectors is 0.

4.2 Planar orthogonality frames

DEFINITION 7.6 *A* 2-dimensional model of orthogonality*, or simply a* planar *model of orthogonality, is an orthogonality frame* $\langle \mathbf{Li}, \perp \rangle$ *with* $\mathbf{Li}$ *non-empty and subject to the following axioms:*

$\mathsf{Irr}_\perp$: $\neg\exists x\,(x \perp x)$
$\mathsf{Sym}_\perp$: $\forall x \forall y\,(x \perp y \to y \perp x)$
$\mathsf{prd}^{(2)}_\perp$: $\forall x \exists y\,(y \perp x)$
$\mathsf{Prd}^{(2)}_\perp$: $\forall x \forall y_1 \forall y_2 \left(\bigwedge_{i=1,2} y_i \perp x \to y_1 \shortparallel y_2\right)$

The axioms $\mathsf{Irr}_\perp$ and $\mathsf{Sym}_\perp$ specify respectively the irreflexivity and symmetry of $\perp$, while $\mathsf{prd}^{(2)}_\perp$ and $\mathsf{Prd}^{(2)}_\perp$ combined state that, up to parallelism, every line has a unique line orthogonal to it. It can easily be verified that the axioms $\mathsf{dim}^{(2)}_\perp$ and $\mathsf{Dim}^{(2)}_\perp$ hold in these structures. It is useful to note that Goldblatt, 1987 studies orthogonality structures which admit *self-orthogonal* lines (lines which lie orthogonal to themselves) as well as *singular* lines (lines which lie orthogonal to all lines in the structure).

We will also make use of the following axiom schemes:

- $\mathsf{Inf}_\infty := \{\lambda_k\}_{k\in\mathbb{N}}$, stating the existence of infinitely many lines, where $\lambda_k := \exists x_1 \ldots \exists x_k\,(\wedge_{i\neq j} x_i \neq x_j)$;
- $\mathsf{Ser}_\perp$, stating that every parallel class has infinite cardinality, consisting of the axioms $\forall x \exists y_1 \ldots \exists y_k \left(\mathrm{Diff}_k(y_1 \ldots y_k) \wedge \wedge_{i=1}^k y_i \shortparallel x\right)$ for every $k \in \mathbb{N}$;
- $\mathsf{Par}_\perp$, stating that there are infinitely many parallel classes, consisting of the axioms $\exists x_1 \ldots \exists x_k\,(\wedge_{i\neq j} x_i \nshortparallel x_j)$ for every $k \in \mathbb{N}$.

Using an approach similar to the proof of Proposition 7.1, we can map lines in an abstract planar orthogonality model to lines in the real plane, to obtain the following.

PROPOSITION 7.7 (REPRESENTATION THEOREM) *Every planar model of line orthogonality with cardinality at most the continuum is isomorphic to a real planar model of line orthogonality.*

By bijectively associating pairs of mutually orthogonal parallel classes in any two countable planar orthogonality models, we furthermore obtain the following.

PROPOSITION 7.8 *The theory of the class of planar orthogonality models satisfying the schemes* $\mathsf{Ser}_\perp$ *and* $\mathsf{Par}_\perp$ *is* ω*-categorical.*

COROLLARY 7.9 *The theory of the class of planar orthogonality models satisfying the schemes* $\mathsf{Ser}_\perp$ *and* $\mathsf{Par}_\perp$ *is complete and decidable.*

4.3 Orthogonality frames in higher dimensions

Given an n-dimensional orthogonality frame, a set of lines $x_1, \ldots, x_k$ in that frame (with $n \geq 1$ and $k \leq n$) will be called *independent* when the formula

$$\text{(7.2)} \qquad \mathrm{LI}_k^{(n)} : \neg\exists z_1 \ldots \exists z_{n-k+1} \left(\bigwedge_{i \neq j} z_i \perp z_j \wedge \bigwedge_{i,j} z_i \perp x_j \right)$$

is satisfied. Lines which are not independent will be called *dependent*. The formula (7.2) takes $k \leq n$ arguments. For $k > n$ define the lines $x_1, \ldots, x_k$ to be dependent and the formula $\mathrm{LI}_k^{(n)}(x_1, \ldots, x_k)$ to be false.

Say that y *lies in the* span *of* $x_1, \ldots, x_k$ when the following formula holds:

$$\text{(7.3)} \qquad \mathrm{Span}(x_1, \ldots, x_k, y) := \forall z \left(\wedge_{i=1}^k z \perp x_i \rightarrow z \perp y \right).$$

Independence and span of lines is an abstraction of the notion of linear independence and span of vectors in a vector space.

DEFINITION 7.10 *An orthogonality frame* $\langle \mathbf{Li}, \perp \rangle$ *with* $\mathbf{Li}$ *non-empty will be called an* n-dimensional model of orthogonality*, where* $n \geq 3$*, if it satisfies the axioms* $\mathsf{Irr}_\perp$, $\mathsf{Sym}_\perp$ *and* $\mathsf{Dim}_\perp^{(n)}$*, together with the axioms*

$$\mathsf{prd}_\perp^{(n)} : \forall x_1 \ldots \forall x_{n-1} \exists y \left(\wedge_{i=1}^{n-1} y \perp x_i \right)$$

$$\mathsf{Prd}_\perp^{(n)} : \forall x_1 \ldots \forall x_{n-1} \Big(\mathrm{LI}_{n-1}^{(n)}(x_1, \ldots, x_{n-1}) \rightarrow \forall y_1 \forall y_2 \left(\wedge_{i,j} y_i \perp x_j \rightarrow y_1 \shortparallel y_2 \right) \Big)$$

From the axiom $\mathsf{prd}_\perp^{(n)}$ it is immediate that the axiom $\mathsf{dim}_\perp^{(n)}$ will hold in all n-dimensional orthogonality models, and in the case where $n = 3$ it can also be shown that the axiom $\mathsf{Dim}_\perp^{(3)}$ may be dropped. The axioms $\mathsf{prd}_\perp^{(n)}$ and $\mathsf{Prd}_\perp^{(n)}$ imply that every n-dimensional orthogonality model has the property

$$\text{(7.4)} \qquad \forall x_1 \ldots \forall x_{n-1} \Big(\mathrm{LI}_{n-1}^{(n)}(x_1, \ldots, x_{n-1}) \rightarrow \exists y \Big(\bigwedge_{i=1}^{n-1} y \perp x_i \wedge \forall z \Big(\bigwedge_{i=1}^{n-1} z \perp x_i \rightarrow z \shortparallel y \Big) \Big) \Big)$$

i.e. every $n-1$ independent lines $x_1, \ldots, x_{n-1}$ have a unique parallel class, which we shall call the *product* of $x_1, \ldots, x_{n-1}$ - denote it as $x_1 \times \cdots \times x_{n-1}$ - that lies orthogonal to all of the x_i. Line products are an abstraction of the vector cross product in $\mathbb{R}^3$. In the context of pencils, $x_1 \times \cdots \times x_{n-1}$ will not be a parallel class, but simply a single line. Note that the operation $\times$ is *not* total, but only defined on tuples of independent lines. In the 3-dimensional case, the formula (7.4) reduces to $\forall x_1 \forall x_2\, (x_1 \not\parallel x_2 \rightarrow \exists y\, (y \perp x_1, x_2 \wedge \forall z\, (z \perp x_1, x_2 \rightarrow z \parallel y)))$, i.e. every two non-parallel lines have a unique parallel class orthogonal to both of them.

The axioms used above for the formalization of orthogonality in higher dimensions illustrate the novel expressive power of orthogonality, but they do not constitute a complete first-order axiomatization of the class of orthogonality structures in dimension $n \geq 3$. To our knowledge, the complete axiomatization of the first-order theory of orthogonality in these dimensions has not been established yet. Unlike the case for planar orthogonality models, it can be shown that the theory of line orthogonality in Euclidean n-space is not countably categorical for $n \geq 3$, and this negative result indicates that the problem of identifying this theory is presumably difficult.

Since orthogonality is a metric notion - arguably the simplest and most intuitive of all metric line notions - it has great expressive power, and one anticipates that its theory will capture a non-trivial and substantial fragment of that of the full Euclidean geometry, as witnessed by the fact that notions like linear independence and span of vectors can be abstracted and expressed in the language of orthogonality.

5. Two-sorted point-line incidence spaces

In this section, we consider point-line incidence structures, described by a two-sorted first-order language with equality, equipped with sorts for points and lines and the intersort relation of incidence.

5.1 Point-line incidence frames

A *point-line incidence frame* is a two-sorted structure $\langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ where $\mathbf{Po}$ and $\mathbf{Li}$ are non-empty sets and $\mathbf{I} \subseteq \mathbf{Po} \times \mathbf{Li}$ is a symmetric *incidence relation* between them. The elements of $\mathbf{Po}$ are called *points*, and the elements of $\mathbf{Li}$ are called *lines*. If the relation $\mathbf{I}$ holds for a point X and a line x then we use expressions like *X is incident with x*, *X lies on x*, *X belongs to x*, *x passes through X*, *x contains X* etc. When $X\mathbf{I}z$ and $Y\mathbf{I}z$ we also say that *the line z* connects *the points X and Y* while *the point X is in the intersection of the lines y and z* will mean that $X\mathbf{I}y$ and $X\mathbf{I}z$.

We say that the lines x and y are *incident*, denoted x **Inc** y, if they are incident with a common point, formally

$$x \,\mathbf{Inc}\, y := \exists Z(Z\mathbf{I}x \wedge Z\mathbf{I}y).$$

Further, we say that the lines x and y are *intersecting*, denoted x **Int** y, if they are incident and different. Formally

$$x \,\mathbf{Int}\, y \;:=\; x \neq y \wedge x \,\mathbf{Inc}\, y.$$

Given an arbitrary incidence frame $\langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$, we also introduce the relation of *collinearity* of three points

$$\mathbf{Col}(XYZ) := \exists x(X\mathbf{I}x \wedge Y\mathbf{I}x \wedge Z\mathbf{I}x)$$

and that of *co-punctuality* of lines

$$\mathbf{Cop}(x_1 \ldots x_n) := \exists X \left(\bigwedge_{i=1}^{n} X\mathbf{I}x_i \right).$$

Thus incidence of lines is a special case of co-punctuality of lines.

5.2 Linear spaces of incidence

Linear spaces (*not* in the sense of vector spaces) are the most general incidence structures which are geometrically meaningful. It is instructive to note that a number of fundamental concepts in vector spaces, such as independence, basis and dimension can be generalized to linear spaces. The following definition reflects Hilbert's axioms for incidence (see Karzel et al., 1973).

DEFINITION 7.11 *A* linear space *(aka* incidence geometry *or* incidence basis *in Mihalek, 1972) is an incidence frame* $\langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ *in which the following axioms hold:*

LS1 *Every two distinct points are incident with a unique common line.*
LS2 *Every line passes through at least two points.*

Given distinct points X and Y in a linear space, the unique line incident with both of them will be denoted by $\mathbf{l}(X, Y)$. Thus, the expression $\mathbf{l}(X, Y)$ assumes that X and Y are distinct. Furthermore, if two lines x and y in a linear space intersect, then by LS1 they have a unique common point, hereafter denoted as $\mathbf{P}(x, y)$ and called the *intersection* of x and y. We will only use this notation in the case of intersecting lines.

5.3 Linear subspaces, independence, bases and dimension

To begin with, linear spaces can be regarded as two-sorted *algebraic* structures, with two partial operations: one applied to two different points produces

the unique line passing through them, and the other applied to two intersecting lines produces their intersection point. Thus, a subspace of a linear space could be defined as a non-empty substructure which is closed under these partial operations. However, this definition also allows subspaces consisting of just one line from the space, and two or more, but not all, points on that line. It is more natural to require that a subspace contains with every line in it all points on that line. Therefore, by a (*linear*) *subspace* (*linear variety* in Gemignani, 1971) of a linear space $\mathfrak{L}$ we will mean every substructure $\mathfrak{L}' = \langle \mathbf{Po}', \mathbf{Li}', \mathbf{I} \rangle$ of the incidence frame $\mathfrak{L}$, which is itself a linear space, and all points lying on lines in $\mathbf{Li}'$ are in $\mathbf{Po}'$. Note that any non-empty intersection of a family of subspaces (i.e. incidence structure in which the sets of lines and points are the respective non-empty intersections of the families of lines and points of the spaces in the family) is a subspace itself. The subspace $\mathfrak{L}'$ of $\mathfrak{L}$ is *generated by the pair of sets of points and lines* $(\mathbf{P}, \mathbf{L})$ in $\mathfrak{L}$, denoted here $\mathfrak{L}' = \lfloor \mathbf{P}, \mathbf{L} \rfloor$, if it is the smallest subspace (i.e., the intersection of all these) of $\mathfrak{L}$ containing the points (lines). Alternatively, $\mathfrak{L}'$ can be obtained from $(\mathbf{P}, \mathbf{L})$ by a finite number of successive steps of adding the line passing through two given points and adding all points lying on a given line. Clearly, it suffices to generate subspaces starting from sets of points only; then we write simply $\mathfrak{L}' = \lfloor \mathbf{P} \rfloor$.

A set of points $\mathbf{P}$ in a linear space is *independent* if none of the points of $\mathbf{P}$ belongs to the subspace generated by the rest of $\mathbf{P}$. An independent set of points which generates (paired with the empty set of lines) a given subspace is called a *basis* of that subspace. The reader is also referred to Ch. 12 for a discussion of matroids and independent sets.

Note that, unlike vector spaces, not all bases in a linear space need to have the same cardinality; for an example, see e.g. Batten, 1986, Sec. 2.1. However, as shown there:

PROPOSITION 7.12 *All bases of a linear space* $\mathfrak{L} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ *have the same cardinality provided that the space satisfies the following* exchange property*:* for any $\mathbf{P} \subset \mathbf{Po}$ and $X, Y \in \mathbf{Po}$, if $X \notin \lfloor \mathbf{P} \rfloor$ and $X \in \lfloor \mathbf{P} \cup \{Y\} \rfloor$ then $Y \in \lfloor \mathbf{P} \cup \{X\} \rfloor$.

The *dimension* of a subspace $\mathfrak{L}'$ of a linear space $\mathfrak{L}$ is the least number n such that $\mathfrak{L}'$ can be generated by a (clearly independent) set of $n + 1$ points. Thus, every subspace containing just one line has dimension 1; a subspace containing three non-collinear points has dimension at least 2. A subspace with dimension 2 of a linear space $\mathfrak{L}$ is a (*linear*) *plane* in $\mathfrak{L}$.

Examples of linear spaces include the usual Euclidean plane and 3D-space, as well as the points in any open disc in the Euclidean plane or space, where lines are the intersections of the usual lines in the plane (space) with that disc. Besides, there is a huge variety of finite linear spaces (see e.g. Batten, 1986, Mihalek, 1972). For instance, take the vertices of a tetrahedron as points, and

the sides of the tetrahedron as lines, where incidence is standard. Note that this is the simplest example of a linear space of dimension greater than 2.

5.4 Linear transformations and collineations

Given linear spaces $\mathfrak{L} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ and $\mathfrak{L}' = \langle \mathbf{Po}', \mathbf{Li}', \mathbf{I}' \rangle$, a mapping f from $\mathfrak{L}$ to $\mathfrak{L}'$ is a pair of mappings $f_{po} : \mathbf{Po} \longrightarrow \mathbf{Po}'$ and $f_{li} : \mathbf{Li} \longrightarrow \mathbf{Li}'$. Such mapping is a *linear transformation* if it preserves incidence both ways, i.e. $X\mathbf{I}x$ iff $f_{po}(X)\mathbf{I}'f_{li}(x)$. Thus, the action of a linear transformation on a line is determined by its action on any two distinct points of the line, and therefore it suffices to consider linear transformations as mappings on the set of points of a linear space. It is immediate from the definition to see that if $\mathfrak{L}$ has dimension greater than 1 then every linear transformation on $\mathfrak{L}$ is injective on the set of points, and on the set of lines, of $\mathfrak{L}$. Thus the notion of linear transformation is the natural notion of a mapping between linear spaces that preserves their structure.

An *isomorphism* between linear spaces is a bijective (on each of the sets of points and lines) linear transformation. A *collineation* is an automorphism of a linear space, i.e. an isomorphism of a linear space onto itself. With ι as the identity and function inverse and composition as basic operations, the set of all collineations of a linear space $\mathfrak{L}$ is a group, called the *group of collineations* $\mathsf{Aut}(\mathfrak{L})$ of $\mathfrak{L}$. Many properties of a linear space can be determined by its group of collineations; for more detail see e.g. Gemignani, 1971, Behnke et al., 1974, Coxeter, 1969, Hughes and Piper, 1973.

5.5 Parallelism and planarity in linear spaces

Given a linear space $\mathfrak{L} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$, one way to define parallel lines in it is to take intuition from the Euclidean space, where two lines are parallel if they are co-planar but do not intersect. Thus, we call two lines in $\mathfrak{L}$ *(strictly) quasi-parallel* if they have no common incident point and belong to a subspace of dimension 2. Note that this relation need not be transitive. For technical reasons, however, we consider separately the case where every line in the space is incident with exactly 2 points. We will call such spaces *meagre*. In meagre spaces, by strict quasi-parallelism we will mean simply non-incidence.

An alternative definition is based on another intuition from "real Euclidean geometry": two lines are parallel if they are not incident, but the diagonals of every quadrilateral with a pair of opposite sides lying on these lines must intersect. Formally, we define the relation $\parallel$ of (*strictly*) *parallel* lines in $\mathfrak{L}$ as follows:

$$\begin{aligned} x \parallel y \;:=\; & \neg\, x \;\mathbf{Inc}\; y \;\wedge \forall X_1 \forall X_2 \forall Y_1 \forall Y_2 \Big(\big(x = \mathbf{l}(X_1, X_2) \wedge y = \mathbf{l}(Y_1, Y_2) \big) \\ & \rightarrow \big(\mathbf{l}(X_1, Y_1) \;\mathbf{Int}\; \mathbf{l}(X_2, Y_2) \;\vee\; \mathbf{l}(X_1, Y_2) \;\mathbf{Int}\; \mathbf{l}(X_2, Y_1) \big) \Big). \end{aligned}$$

Then we define *weak parallelism*:

$$x \shortparallel y \;:=\; x \parallel y \vee x = y.$$

Again, in the special case of meagre linear spaces, by strict parallelism we mean non-incidence.

The relation $\parallel$ is irreflexive and symmetric but not necessarily transitive either. Still, if two lines in a linear space are parallel, then they are quasi-parallel too: take any two pairs of distinct points, one on each of the lines, and take the intersection point of the respective "diagonals"; that point together with the pair of points on any of the lines generates a subspace of dimension 2 containing both lines. The converse need not hold, which can be shown by an example from Batten, 1986, Sec. 2.1, of a linear space of dimension 2 which contains a set of 4 independent points, mentioned in Sec. 5.3.

Given a linear space $\mathfrak{L} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$, two lines $x, y \in \mathbf{Li}$ are called *co-planar*, denoted $\mathbf{Pl}(xy)$, if they are incident or parallel:

$$\mathbf{Pl}(xy) \;:=\; x \;\mathbf{Inc}\; y \vee x \parallel y.$$

A linear space is *planar* if every two lines in it are co-planar.

Thus, by convention, every meagre linear space is planar.

It is easy to see that a non-meagre linear space $\mathfrak{L}$ is planar iff it satisfies the following property:

$$\begin{aligned} &\forall X_1 \forall X_2 \forall X_3 \forall X_4 \Big(\mathbf{l}(X_1, X_2) \;\mathbf{Inc}\; \mathbf{l}(X_3, X_4) \;\vee \\ &\quad \mathbf{l}(X_1, X_3) \;\mathbf{Inc}\; \mathbf{l}(X_2, X_4) \;\vee\; \mathbf{l}(X_1, X_4) \;\mathbf{Inc}\; \mathbf{l}(X_2, X_3) \Big), \end{aligned} \tag{7.5}$$

saying that the diagonals of every quadrilateral must intersect.

We now have two different notions of a "plane" in a linear space, one based on dimension, the other on planarity. Note that the tetrahedron is a meagre (and hence planar) space of dimension 3. However, if a non-meagre linear space $\mathfrak{L}$ is planar, then it has dimension 2. Indeed, let A, B, C be any three non-collinear points in it. Then for any point D in the space, at least one of the pairs of lines $(\mathbf{l}(A, B), \mathbf{l}(C, D))$, $(\mathbf{l}(A, C), \mathbf{l}(B, D))$ and $(\mathbf{l}(A, D), \mathbf{l}(B, C))$ are incident, say $\mathbf{l}(A, B) \;\mathbf{Inc}\; \mathbf{l}(C, D)$, and let $X = \mathbf{P}(\mathbf{l}(A, B), \mathbf{l}(C, D))$. Then X belongs to the line $\mathbf{l}(A, B)$ and D belongs to the line $\mathbf{l}(C, X)$.

The converse of the claim above need not hold, again by the example from Batten, 1986 mentioned above; note that any independent set of 4 points in a non-meagre space violates the planarity condition above.

5.6 Projective spaces and planes

DEFINITION 7.13 *A* projective space *is linear space satisfying the following additional axioms:*

PS1 *If A, B, C are distinct points and a line l intersects AB and AC in two distinct points, then it intersects BC as well.*
PS2 *There are at least four points, no three of which are collinear.*

(See also Lenz, 1954.)

Projective spaces satisfy the exchange property (see Batten, 1986, Sec. 3.9), and hence, by Proposition 7.12, every two bases in a projective space have the same cardinality, called the *rank* of the space. The dimension of the space is thus defined as 1 less than its rank.

A *projective plane* is a projective space of rank 3, i.e. dimension 2. Equivalently, a projective plane is a projective space in which every two lines intersect; in particular, projective planes contain no parallel lines.

Conversely, one can re-define projective spaces of higher dimension in terms of the sub-planes that they contain. For instance, the *projective 3D-space* can be defined (see Hartshorne, 1967, Mihalek, 1972) as a projective space with the following additional axioms:

PS3 *There exist at least 4 non-coplanar points.*
PS4 *Every three non-collinear points lie on a unique sub-plane.*
PS5 *Every line meets every sub-plane in at least one point.*
PS6 *Every two sub-planes have at least one common line.*

Since every line in a projective plane intersects all other lines in different points, and every point is line-connected with every other point in the plane, it follows that in a finite projective plane every line is incident with the same number of points, and every point is incident with the same number of lines.

If φ is any statement about a projective plane formulated in terms of "point", "line" and "incidence" then the statement φ^*, formed from φ by interchanging the words "point" and "line", is called the *dual statement* (with respect to "point" and "line") of φ. A statement φ is *self-dual* if $\varphi = \varphi^*$. A theorem about projective planes formulated in terms of the notions "point", "line" and "incidence" is a *projective validity* if it is true in the class of all projective planes, i.e. it is derivable from the axioms for projective planes. One reason why projective planes are interesting is the following "two for the price of one" result (see e.g., Mihalek, 1972, Hughes and Piper, 1973, Batten, 1986):

THEOREM 7.14 (DUALITY PRINCIPLE FOR PROJECTIVE PLANES) *Let φ be a projective validity. Then the dual φ^* of φ is also a projective validity.*

To prove the duality principle it suffices to note that the duals of the axioms for projective planes provide an equivalent axiomatization, and hence the "dual" of every proof in the (first-order) theory of projective planes is a proof in that theory, too. For instance, it follows from the duality principle that in every projective plane there are at least four lines, no three of them incident with the

same point. The reader is also referred to the self-dual axiomatizations of Esser, 1951, Esser, 1973, Kordos, 1982, Menger, 1948, Menger, 1950.

To illustrate the power of the duality principle, consider the following combinatorial example. Suppose we have some projective plane with the property that every line contains n points. Now fix any line l and any point P not on l. Then every point X on l determines a line PX and since l contains n points then there must be n distinct lines of the form PX. Furthermore, note every point in the plane must lie on exactly one of these lines PX. If all these lines were disjoint (as sets of points) then there would be n^2 points in total, but since the point P is counted n times then the total number of points in the plane is $n^2 - (n-1) = n^2 - n + 1$. Thus, we have just shown that the following is a projective validity:

- If every line is incident with n points, then there are $n^2 - n + 1$ points in the entire plane.

By the duality principle, we can conclude the dual of this result:

- If every point is incident with n lines, then there are $n^2 - n + 1$ lines in the entire plane.

Finally, note that axiom PS2 implies that every line in a projective plane is incident with at least 3 points, and therefore, by the result above, the least projective plane, known as *Fano plane*, given on Fig. 7.1, has 7 points and 7 lines.

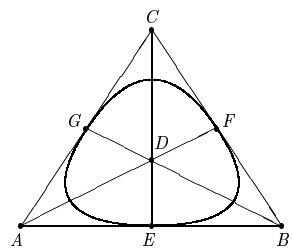


Figure 7.1.

5.7 Affine spaces and planes

DEFINITION 7.15 *An* affine structure *is a linear space* $\mathfrak{L} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ *in which the relation of weak parallelism* $\parallel$ *is an equivalence relation.*

An affine space *is an affine structure with at least 3 non-collinear points, in which the following axiom (Euclid's Fifth postulate) holds:*

V : *Given a line x and a point X not on x, there is a unique line through X that is (strictly) parallel to x, denoted by* $\mathbf{p}(X, x)$.

(See also Lenz, 1954, Lenz, 1989.)

A slightly more general, but essentially equivalent, definition is:

V' : Given a line x and a point X, there is a unique line through X that is weakly parallel to x, denoted by $\mathbf{p}(X, x)$.

An (*affine*) *subspace* of an affine space $\mathfrak{L}$ is any linear subspace of $\mathfrak{L}$ containing 3 non-collinear points and satisfying V itself, i.e. closed under the operation $\mathbf{p}$.

Examples of affine spaces include the usual Euclidean plane and 3D-space, but not the open disc in the Euclidean plane or space, as the axiom V fails there. An example of a finite affine space is the tetrahedron. For other examples, see e.g. Batten, 1986 and Coxeter, 1969.

Note that, usually, the literature on affine and projective geometry deals only with *affine planes*, and only occasionally introduces higher-dimensional affine spaces. Thus, our definition is somewhat more general, as it is based neither on coordinatization of the space, nor on the earlier-defined notion of affine plane. A small price to pay for that generality was the adjustment of the definitions of parallelism and planarity in the special case of meagre spaces.

Planes and planarity can be re-defined in non-meagre affine spaces using the following observations. Any "triangle" (three non-collinear points) together with the three lines determined by these points, must define a plane. By the Fifth Postulate, every line in that plane must be incident with at least two of these three lines, i.e. every line is determined by a pair of different points, each incident with some of these lines. Furthermore, every point in that plane belongs to at least one line constructed in this way, e.g. any line determined by that point and a vertex of the original triangle, which intersects the side opposite to that vertex (there will be at least one, by planarity). Turning these observation around, we obtain the following definition: a *plane in an affine space* $\mathfrak{L}$ is an incidence structure $\mathfrak{S}$ constructed as follows: take three non-collinear points P_1, P_2, P_3 in $\mathfrak{L}$ and the lines in $\mathfrak{L}$ determined by them, say x_1, x_2, x_3, where $x_i = \mathbf{l}(P_j, P_k)$, for i, j, k pairwise different. Let $\mathbf{P}$ be the set of points in $\mathfrak{L}$ incident with at least one of the lines x_1, x_2, x_3. Then the lines in $\mathfrak{S}$ are exactly those lines in $\mathfrak{L}$ incident with at least two different points from $\mathbf{P}$, and the points in $\mathfrak{S}$ are those points in $\mathfrak{L}$ incident with at least one of these lines. It is not difficult to see that every such plane is an affine subspace of $\mathfrak{L}$, which will be called the (*affine*) *plane in the space* $\mathfrak{L}$ generated by the points P_1, P_2, P_3. In particular, every plane in an affine space is closed under the *affine operations* of taking lines through two points, intersections of lines, and construction of lines passing through a given point and parallel to a given line, based on the axiom V.

Now, an affine space is called *planar*, or an *affine plane*, if it coincides with some plane in it. It is easy to show (see e.g. Batten, 1986) that every affine space satisfies the exchange property, and hence all bases in an affine space have the

same cardinality. Rank and dimension are introduced as in projective spaces. As for the tetrahedron, despite being of dimension 3, we have a good excuse (to become clear further) to consider it an affine plane as well, and that is the main reason to adjust the definition of parallelism and planarity for meagre spaces. In fact (see Batten, 1986, Sec. 4.1), it is the *only* affine plane of dimension more than 2. Moreover, as shown there, every line in a finite affine space is incident with the same number of points, and vice versa.

Thus, if an affine space is planar then the relation of non-incidence of lines is pseudo-transitive, hence it is a relation of strict parallelism. To summarize:

PROPOSITION 7.16 *A non-meagre affine space $\mathfrak{L}$ is an affine plane iff it has a dimension 2 iff the relations of non-incidence and strict parallelism between lines in $\mathfrak{L}$ coincide.*

5.8 Relationship between affine and projective planes

Affine planes do not have the duality property. For example, the dual of LS1 does not hold as it violates Euclid's Parallel Postulate V. Still, there is an intimate relationship between affine planes and projective planes, given by the following two theorems, the proofs of which can be found e.g. in Gemignani, 1971, Hughes and Piper, 1973, Mihalek, 1972.

THEOREM 7.17 *Let $\mathfrak{P} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ be a projective plane and let l^* be a line in $\mathbf{Li}$. Define $\mathbf{Po}^- := \{X \in \mathbf{Po} : X \notin l^*\}$, $\mathbf{Li}^- := \mathbf{Li} \backslash \{l^*\}$ and $\mathbf{I}^- := \mathbf{I}|_{\mathbf{Po}^- \times \mathbf{Li}^-}$. Then the structure $\mathfrak{P}^- = \langle \mathbf{Po}^-, \mathbf{Li}^-, \mathbf{I}^- \rangle$ is an affine plane, called the (*deletion*) affine subgeometry of $\mathfrak{P}$ induced by l^*.*

THEOREM 7.18 *Let $\mathfrak{A} = \langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ be any affine plane and let l^* be any set, disjoint with $\mathbf{Po}$ and $\mathbf{Li}$, and of the same cardinality as the number of parallel classes in $\mathfrak{A}$. To every parallel class $[l]_{\parallel}$ in $\mathfrak{A}$, assign some distinct element $P_{[l]_{\parallel}} \in l^*$ to $[l]_{\parallel}$. Define*

1. $\mathbf{Po}^+ := \mathbf{Po} \cup l^*$;
2. $\mathbf{Li}^+ := \mathbf{Li} \cup \{l^*\}$;
3. $\mathbf{I}^+ := \mathbf{I} \cup \{(P_{[l]_{\parallel}}, l) : l \neq l^*\} \cup \{(P, l^*) : P \in l^*\}$.

Then the structure $\mathfrak{A}^+ = \langle \mathbf{Po}^+, \mathbf{Li}^+, \mathbf{I}^+ \rangle$, which we will call the projective extension *of $\mathfrak{A}$, is a projective plane.*

Thus, affine and projective planes are separated by a single line, the so-called "line at infinity". The tetrahedron and Fano plane (Fig. 7.1) illustrate the latter two results. It is because of the Fano plane that we insisted the tetrahedron, being its deletion subgeometry, should be an affine plane.

Note that the constructions between affine and projective planes described above are mutually inverse, up to isomorphism. These constructions can be described in logical terms, relating the first-order theories of the affine and projective planes. On the one hand, every affine plane is first-order interpretable into its projective extension in an obvious way; on the other hand, the first-order theory of a projective plane can be reduced to the first-order theory of its affine subgeometry. Consequently, given a class of projective planes, its elementary theory is decidable iff the elementary theory of the class of respective affine subgeometries is decidable, too. Likewise, the elementary theory a class of affine planes is decidable iff the elementary theory of the class of respective projective extensions is decidable, too.

6. Coordinatization

In this section we give an overview on the coordinatization and subsequent algebraization of affine planes. We will introduce a special class of algebraic structures called "ternary rings", the elements of which can serve as coordinates of points in the plane. It will turn out that affine planes and ternary rings are inter-definable in the sense that from every affine plane one can extract a ternary ring while every ternary ring gives rise to an affine plane. In fact, these constructions are essentially *logical (first-order) interpretations*, which thus relate their first-order theories. In particular, we will see that natural and important geometric properties of affine planes, viz. Desargues' and Pappus' properties, correspond to natural algebraic properties in these ternary rings. Furthermore, we will demonstrate the interaction between special dilations of affine planes and the properties of Desargues and Pappus, and will discuss the logical consequences of the coordinatization. In particular, we will extract the axiomatizations and (un)decidability of the first-order theories of some important affine planes and classes of planes from their associated coordinate rings.

The method of coordinatization applies likewise to projective planes, and most of the results obtained below have close projective analogues. Since both constructions are very similar, we will only present here coordinatization of affine planes. For a more detailed account of the coordinatization of affine and projective planes and the relationships (with proofs) between geometric and algebraic properties, the reader is referred to Blumenthal, 1961, Artin, 1957, Heyting, 1963, Szmielew, 1983, Mihalek, 1972, Hughes and Piper, 1973, etc.

6.1 Coordinate systems in affine planes

We adopt weak parallelism in the discussion for the rest of this section.

Let any affine plane $\mathfrak{A} = \langle \mathbf{Po}; \mathbf{Li}; \mathbf{I} \rangle$ be given. The following procedure assigns coordinates to the plane.

- Take any triplet of non-collinear points O, X and Y. The point O will be called the *origin* and the triplet OXY will be called the *coordinate system*.
- Let I be the point of intersection of the line in the parallel class of OX containing Y, with the line in the parallel class of OY containing X. The point I will be called the *unit point* while the lines OX, OY and OI will be called respectively the *x-axis*, *y-axis* and *unit line*.
- Let Γ be any abstract set, containing elements 0 and 1, of the same cardinality as the number of points on the unit line. In fact, since all lines in an affine plane contain the same number of points, Γ can have the cardinality of any line in the plane. We call the set Γ the *coordinate set*.
- Now let γ be any bijection between points on the unit line and Γ and such that $\gamma(O) = 0$ and $\gamma(I) = 1$.

Points in the plane are assigned coordinates consisting of ordered pairs from Γ^2 in the following manner:

- If P is a point on the unit line and $\gamma(P) = p$ then the coordinates of P are (p, p).
- Let P be any point not on the unit line. Suppose the line in the parallel class of the y-axis, containing P, intersects the unit line in the point with coordinates (a, a), and suppose that the line in the parallel class of the x-axis, containing P, intersects the unit line in the point with coordinates (b, b). Then the coordinates of P are (a, b) (refer to Fig. 7.2).

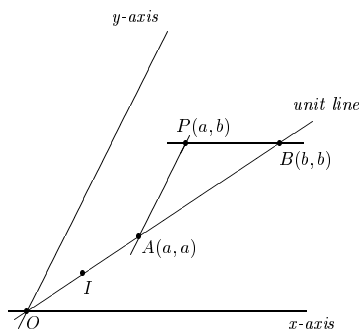


Figure 7.2.

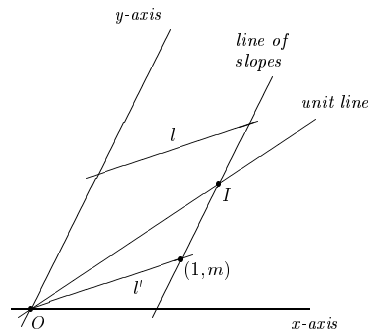


Figure 7.3.

The point P with coordinates (x, y) will be denoted as $P(x, y)$, and points will be identified with their coordinates ($\mathbf{P}(x, y)$ may also refer to the point

of intersection of lines x and y, but the context should make it clear what the intended meaning is). For example, the points O, X, Y and I can be given simply as $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$. The value x is called the *abscissa* of P and y is called the *ordinate* of P. Any line not in the parallel class of the y-axis will intersect the y-axis in some point $(0,c)$. This value c is called the y-*intercept* of the line.

Next we define the *slope* of a line, using what will be called the *line of slopes*, which is that line in the parallel class of the y-axis intersecting the unit point I. Every point on the line of slopes will have coordinates $(1,m)$ for some $m \in \Gamma$. There are two types of line to consider:

1. If l is a line parallel to the y-axis, its slope is left undefined.

2. Consider any line l not in the parallel class of the y-axis. There will be a unique line l' parallel to l and incident with O. This line l' will intersect the line of slopes in some point $(1,m)$. The slope of l is defined as the value m (refer to Fig. 7.3).

Thus the slope of the x-axis is 0 and the slope of the unit line is 1.

The *equation* of a line is any equation formulated in terms of variables x and y such that all and only those points (x,y) belonging to the line satisfy the equation. For example, the line parallel to the x-axis with y-intercept b has equation $y = b$, and the line parallel to the y-axis intersecting the x-axis in the point $(a,0)$ has equation $x = a$. The unit line has equation $y = x$. But we need further algebraic machinery to describe the equations of lines other than these trivial examples. This is provided by the operation $T : \Gamma^3 \to \Gamma$, defined as follows. Let the triple of values $(m,a,c) \in \Gamma^3$ be given. To compute the value of $T(m,a,c)$ consider the line l of slope m and y-intercept c. Then l intersects the line parallel to the y-axis, and containing the point $(a,0)$, in the point $(a, T(m,a,c))$ (refer to Fig. 7.4).

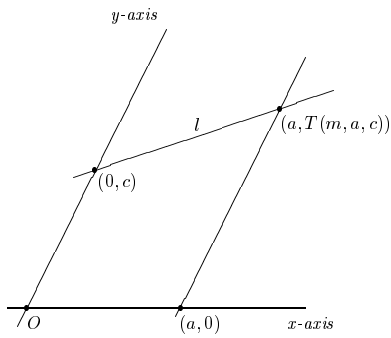


Figure 7.4.

T enables us to obtain an equation for any line in the plane. If the line has undefined slope, it has an equation $x = a$ for some $a \in \Gamma$, while the line with slope m and y-intercept c has equation $y = T(m, x, c)$. The above definition of T uses a geometric construction, but T can also be described purely algebraically using the structure of ternary ring.

6.2 Ternary rings and coordinate systems

DEFINITION 7.19 *A* ternary ring *(also known as a* Hall planar ternary ring*, in reference to Marshall Hall Jr.) is an algebraic structure* $\mathfrak{T} = (F; 0, 1, T)$ *consisting of a set* F *together with distinguished elements* $0, 1 \in F$ *and a ternary operation* T *on* F *subject to the following axioms:*

T_0 : $0 \neq 1$

T_1 : $T(a, 1, 0) = a$

T_2 : $T(1, b, 0) = b$

T_3 : $T(a, 0, c) = c$

T_4 : $T(0, b, c) = c$

T_5 : $T(a, b, x) = d$ *has a solution for* x

T_6 : *If* $T(a, b, c) = T(a, b, c')$ *then* $c = c'$

T_7 : *If* $b \neq b'$ *then the simultaneous equations* $T(x, b, y) = d$ *and* $T(x, b', y) = d'$ *have a solution for* x *and* y.

T_8 : *If* $a \neq a'$ *then* $T(a, x, c) = T(a', x, c')$ *has a solution for* x

T_9 : *For* $b \neq b'$*, if* $T(a, b, c) = T(a', b, c')$ *and* $T(a, b', c) = T(a', b', c')$ *then* $a = a'$ *and* $c = c'$

It is easy to see that the ternary operation T satisfies all the axioms $\mathsf{T}_0 - \mathsf{T}_9$ hence we have the following important result.

THEOREM 7.20 *Let the affine plane* $\mathfrak{A}$ *be coordinatized with coordinate system* OXY *and coordinate set* Γ *and let* T *be the resulting ternary operation on* Γ*. Then the structure* $(\Gamma; 0, 1, T)$ *is a ternary ring.*

The ternary ring $(\Gamma; 0, 1, T)$ above will be called a (*coordinate*) *ring attached to the plane* $\mathfrak{A}$ (by means of the coordinate system OXY), denoted $\mathbf{T}_{OXY}(\mathfrak{A})$. Given the coordinate system OXY, there is only one, up to isomorphism, ternary ring attached to the plane by means of OXY. A ternary ring $\mathfrak{T}$ is said to be *attached* to the affine plane $\mathfrak{A}$ provided there is *some* coordinate system OXY such that $\mathfrak{T} \cong \mathbf{T}_{OXY}(\mathfrak{A})$.

A converse to the last theorem is also true.

THEOREM 7.21 *For any ternary ring $\mathfrak{T} = (F; 0, 1, T)$, the plane $\mathbf{A}(\mathfrak{T})$ with point universe F^2, and line universe consisting of all sets of the form $\{(a, y) : y \in F\}$ and $\{(x, T(m, x, c)) : x \in F\}$ for every $a, m, c \in F$, is an affine plane, called* the affine plane over the ternary ring $\mathfrak{T}$.

The constructions given in the two theorems above are inverse in the following sense. Let an affine plane $\mathfrak{A}$ be given and fix some coordinate system OXY. Then $\mathbf{A}(\mathbf{T}_{OXY}(\mathfrak{A})) \cong \mathfrak{A}$. Let a ternary ring $\mathfrak{T} = (F; 0, 1, T)$ be given. Then $\mathbf{T}_{(0,0)(0,1)(1,0)}(\mathbf{A}(\mathfrak{T})) \cong \mathfrak{T}$.

If two ternary rings are isomorphic then so will be the affine planes over those ternary rings. But surprisingly, there are non-isomorphic ternary rings such that the affine planes over those ternary rings are still isomorphic. In particular, coordinatizing an affine plane with different coordinate systems may sometimes give rise to non-isomorphic ternary rings attached to the same plane, and later on we will give a sufficient condition for uniqueness of the coordinate rings.

Given a ternary ring $(F; 0, 1, T)$, *addition* $+$ and *multiplication* $\cdot$ are defined on F as follows:

$$a + b = T(1, a, b); \quad a \cdot b = T(a, b, 0).$$

The structure $(F; 0, 1, +, \cdot)$ thus formed is actually a double loop. Hence the class of double loops contains the class of ternary rings.

The geometric analogue to addition is the translation of a line in the plane, that of multiplication is the rotation of a line in the plane. To calculate $a + b$ proceed as follows. Take the points $A(a, a)$ and $B(b, b)$ that lie on the unit line. Intersect the line parallel to the x-axis containing B with the y-axis to obtain the point $Q(0, b)$. Then take the line parallel to the unit line containing the point Q and intersect it with the line parallel to the y-axis containing the point A to obtain the point $P(a, c)$. The line parallel to the x-axis containing P intersects the unit line in the point $C(c, c)$. We define $C = A + B$ and $c = a + b$ (refer to Fig. 7.5). To calculate $a \cdot b$ proceed as follows. Take the points $A(a, a)$ and $B(b, b)$ that lie on the unit line. Intersect the line parallel to the x-axis containing A with the line of slopes to obtain the point $Q(1, a)$. Then the line parallel to the y-axis containing the point B will intersect the line OQ in some point P. Intersect the line parallel to the x-axis containing P with the unit line to obtain the point $C(c, c)$. We define $C = A \cdot B$ and $c = a \cdot b$ (refer to Fig. 7.6).

We are now able to give the linear equations of two more classes of lines. The line of slope 1 with y-intercept c will have equation $y = x + c$ while the line of slope m and y-intercept 0 will have equation $y = m \cdot x$.

Call a left division ring *strong* if it satisfies the additional property

$$m_1 \neq m_2 \Rightarrow \forall c_1 \forall c_2 \exists x \, (m_1 \cdot x + c_1 = m_2 \cdot x + c_2) \, ;$$

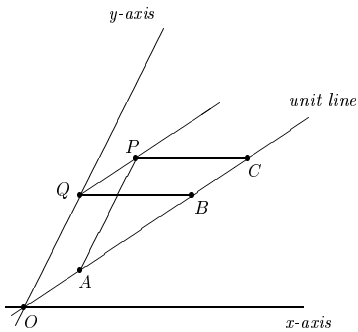


Figure 7.5.

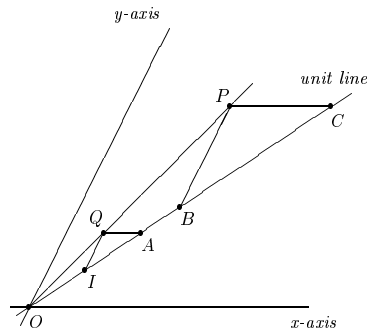


Figure 7.6.

informally, lines with different directions intersect. We can define a ternary operation T in the strong left division ring $(F; 0, 1, +, \cdot)$ as $T(a, b, c) := a \cdot b + c$. Then it turns out that $(F; 0, 1, T)$ will be a ternary ring. Hence the class of ternary rings contains the class of strong left division rings. It is easy to see that every division ring (with full distributivity) is strong. Let $\mathcal{T}$, $\mathcal{LD}$, $\mathcal{D}$, $\mathcal{SF}$ and $\mathcal{F}$ denote respectively the classes ternary rings, strong left division rings, division rings, skew fields and fields. Then we have the following chain if inclusions:

$$\mathcal{T} \supseteq \mathcal{LD} \supseteq \mathcal{D} \supseteq \mathcal{SF} \supseteq \mathcal{F}.$$

6.3 The properties of Desargues and Pappus

The theorems of Desargues and Pappus, known from Euclidean geometry, turn out to hold in a more general, affine setting. While retaining their elementary nature, these properties of the Euclidean plane will lose their status as theorems when taken in arbitrary affine planes, as they may or may not hold true depending on the affine plane concerned. The Desargues and Pappus properties deal with configurations of six points, in the former case lying in pairs on three lines, and in the latter case lying in triples on two lines. We distinguish cases where the lines (i) are parallel, or (ii) are mutually incident. The specific geometric properties thus described will correspond to specific classes of the algebraic structures described above.

DEFINITION 7.22 *An affine plane satisfies the* First Desargues Property D_1, *if the following holds (see Fig. 7.7)*:

$$\begin{aligned}\mathsf{D}_1 :\ \big(\neg\mathbf{Col}(AA'B) \wedge \neg\mathbf{Col}(AA'C) \wedge AA' \parallel BB' \wedge AA' \parallel CC' \\ \wedge\ AB \parallel A'B' \wedge AC \parallel A'C'\big) \rightarrow\ BC \parallel B'C'\end{aligned}$$

The Euclidean plane satisfies D_1. As an example (see Blumenthal, 1961) of a plane that does not satisfy D_1, consider the real plane $\mathbb{R}^2$ with lines modified

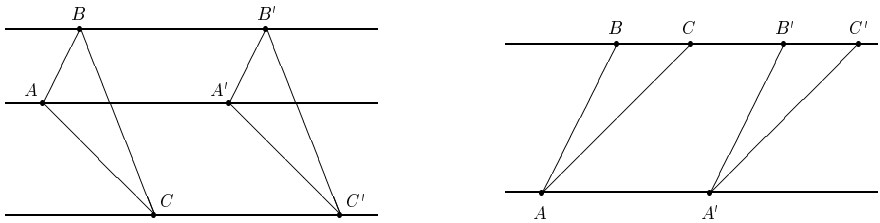


Figure 7.7.

as follows. Any line of either undefined slope or non-positive slope will be left unaltered, but any line $y = mx + c$ with strictly positive slope $m > 0$ is changed to the union of the two rays

$$\begin{aligned} y &= mx + c \quad \text{when} \quad y < 0, \\ y &= \tfrac{1}{2}mx + \tfrac{1}{2}c \quad \text{when} \quad y \geq 0. \end{aligned}$$

It is easy to see that this modified plane is affine. Fig. 7.8 gives a configuration of points that falsify D_1.

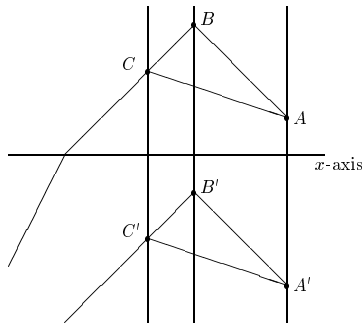


Figure 7.8.

We have the following representation theorem.

THEOREM 7.23 *If an affine plane $\mathfrak{A}$ satisfies D_1 then $\mathfrak{A} \cong \mathbf{A}(\mathfrak{F})$ for some strong left division ring $\mathfrak{F} \in \mathcal{LD}$. Conversely, if $\mathfrak{F} \in \mathcal{LD}$ is any strong left division ring, then $\mathbf{A}(\mathfrak{F})$ satisfies D_1.*

The First Desargues Property allows us to give the linear equation for any line with slope m and y-intercept c. Say that a ternary ring $(R; 0, 1, T)$ is *linear* if $T(a, b, c) = a \cdot b + c$ for all $a, b, c \in R$ (where $+$ and $\cdot$ are interpreted in the expanded structure $(R; 0, 1, +, \cdot, T)$).

THEOREM 7.24 *An affine plane $\mathfrak{A}$ satisfies the property D_1 if and only if every ternary ring $(\Gamma; 0, 1, T)$ attached to $\mathfrak{A}$ is linear.*

If $\mathfrak{A}$ is an affine plane satisfying D_1, then every line of undefined slope has the form $x = a$ while the line of slope m with y-intercept c has equation $y = m \cdot x + c$. This concludes the task of finding a linear equation for every line of the plane.

Let $\mathbf{P}$ be the quaternary *parallelogram relation* defined by

$$\mathbf{P}(ABCD) \Leftrightarrow AB \parallel CD \wedge AC \parallel BD.$$

The property D_1 guarantees that $\mathbf{P}$ will be transitive in the sense

$$\mathbf{P}(ABCD) \wedge \mathbf{P}(ABEF) \Rightarrow \mathbf{P}(CDEF)$$

where the points are so as to exclude obvious degenerate cases.

DEFINITION 7.25 *An affine plane is said to satisfy the* Second Desargues Property D_2 *if the following holds (see Fig. 7.9)*:

$$\begin{aligned}\mathsf{D}_2 : \ \big(&\mathrm{Diff}_7(OABCA'B'C') \wedge \neg\mathbf{Col}(ABC) \wedge \neg\mathbf{Col}(A'B'C') \\ &\wedge \mathbf{Col}(OAA') \wedge \mathbf{Col}(OBB') \wedge \mathbf{Col}(OCC') \wedge BC \parallel B'C' \\ &\wedge AC \parallel A'C' \wedge AC \parallel OB\big) \rightarrow \ AB \parallel A'B' \ .\end{aligned}$$

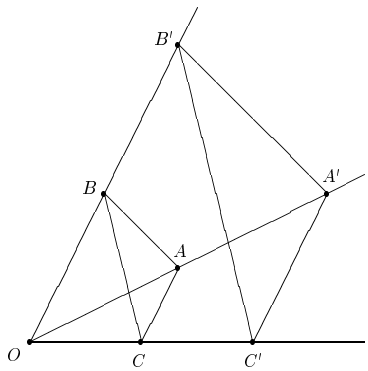


Figure 7.9.

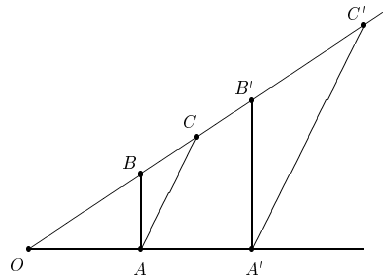


Figure 7.10.

D_2 (also known in the German-language literature as *Trapezdesargues*) endows the attached ternary ring with right distributivity.

THEOREM 7.26 *Let an affine plane $\mathfrak{A}$ satisfy both the properties D_1 and D_2. Then every ternary ring attached to $\mathfrak{A}$ is a division ring.*

DEFINITION 7.27 *An affine plane is said to satisfy the* Third Desargues Property D_3 *if the following holds (see Fig. 7.9 and Fig. 7.10):*

$$\begin{aligned}\mathsf{D}_3 : \ \big(&(O \neq A, B, C, A', B', C') \wedge \neg\mathbf{Col}(AA'B) \wedge \neg\mathbf{Col}(AA'C) \\ &\wedge \mathbf{Col}(OAA') \wedge \mathbf{Col}(OBB') \wedge \mathbf{Col}(OCC') \\ &\wedge AB \parallel A'B' \wedge AC \parallel A'C'\big) \rightarrow \ BC \parallel B'C' \ .\end{aligned}$$

Clearly $\mathsf{D}_3 \Rightarrow \mathsf{D}_2$ and it can also be shown that $\mathsf{D}_3 \Rightarrow \mathsf{D}_1$.

D_3 endows the ternary ring attached to a plane with associative multiplication. We have the following representation theorem.

THEOREM 7.28 *If an affine plane $\mathfrak{A}$ satisfies D_3 then $\mathfrak{A} \cong \mathbf{A}(\mathfrak{F})$ for some skew field $\mathfrak{F} \in \mathcal{SF}$. Conversely, if $\mathfrak{F} \in \mathcal{SF}$ is any skew field, then $\mathbf{A}(\mathfrak{F})$ satisfies D_3.*

For any point O, let $\mathbf{T}_O$ be the quaternary *trapezium relation*:

$$\mathbf{T}_O(ABCD) \Leftrightarrow \mathbf{Col}(OAB) \wedge \mathbf{Col}(OCD) \wedge AC \parallel BD.$$

The property D_3 guarantees that $\mathbf{T}_O$ will be transitive in the sense

$$\mathbf{T}_O(ABCD) \wedge \mathbf{T}_O(ABEF) \Rightarrow \mathbf{T}_O(CDEF)$$

where the points are taken so as to exclude obvious degenerate cases.

In particular, note that in any affine plane satisfying the Third Desargues Property the midpoint of a line segment AB, being the intersection point of the diagonals AB and CD of any parallelogram $ACBD$, is definable in terms of A and B.

As shown in Szmielew, 1983, if the plane satisfies D_3, then the coordinate ternary ring attached to the plane is invariant, up to isomorphism, of the coordinate system used; equivalently, every skew field can be restored uniquely from the affine plane over it. Formally:

THEOREM 7.29 (SZMIELEW, 1983, SEC. 4.5) *If $\mathfrak{F}, \mathfrak{T}$ are skew fields such that $\mathbf{A}(\mathfrak{F}) \cong \mathbf{A}(\mathfrak{T})$ then $\mathfrak{F} \cong \mathfrak{T}$.*

COROLLARY 7.30 *If $\mathfrak{A}$ satisfies D_3 and OXY, $O'X'Y'$ are two coordinate systems in $\mathfrak{A}$ then $\mathbf{T}_{OXY}(\mathfrak{A}) \cong \mathbf{T}_{O'X'Y'}(\mathfrak{A})$.*

Hereafter, whenever $\mathfrak{A}$ satisfies D_3 we will denote the unique coordinate ring attached to $\mathfrak{A}$ by $\mathbf{T}(\mathfrak{A})$.

DEFINITION 7.31 *An affine plane satisfies the* First Pappus Property P_1 *if the following holds (see Fig. 7.11)*:

$$\begin{aligned}\mathsf{P}_1 :\ \big(&\mathbf{Col}(ABC) \wedge \mathbf{Col}(A'B'C') \wedge AB \parallel A'B' \\ &\wedge AB' \parallel A'B \wedge AC' \parallel A'C\big) \rightarrow\ BC' \parallel B'C.\end{aligned}$$

DEFINITION 7.32 *An affine plane satisfies the* Second Pappus Property P_2 *if the following holds (see Fig. 7.12)*:

$$\begin{aligned}\mathsf{P}_2 :\ \big(&(O \neq A, B, C, A', B', C') \wedge \mathbf{Col}(OABC) \wedge \mathbf{Col}(OA'B'C') \\ &\wedge AB \neq A'B' \wedge AB' \parallel A'B \wedge AC' \parallel A'C\big) \rightarrow\ BC' \parallel B'C.\end{aligned}$$

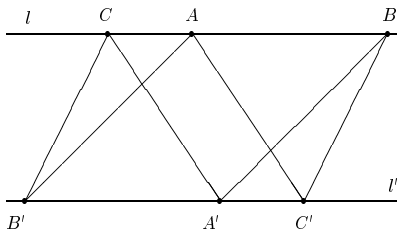


Figure 7.11.

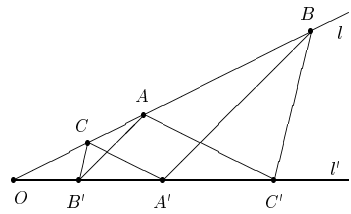


Figure 7.12.

It can be shown that $\mathsf{P}_2 \Rightarrow \mathsf{P}_1$. A famous theorem by Hessenberg establishes the implication $\mathsf{P}_2 \Rightarrow \mathsf{D}_3$. In fact, the following string of implications holds:

$$\mathsf{P}_2 \Rightarrow \mathsf{D}_3 \Rightarrow \mathsf{D}_1 \Rightarrow \mathsf{P}_1.$$

From algebraic considerations one can also derive

$$\mathsf{D}_1 \not\Rightarrow \mathsf{D}_3 \not\Rightarrow \mathsf{P}_2.$$

For instance, the affine plane over the skew field of quaternions satisfies D_3, but not P_2, because of Theorem 7.33 below. It remains an open problem whether P_1 implies (and hence is equivalent to) D_1. See Szmielew, 1983 and Menghini, 1991 for further details.

The property P_2 endows the attached ternary ring with commutative multiplication, and the following representation theorem holds.

THEOREM 7.33 *If an affine plane $\mathfrak{A}$ satisfies P_2 then $\mathfrak{A} \cong \mathbf{A}(\mathfrak{F})$ for some field $\mathfrak{F}$. Conversely, if $\mathfrak{F}$ is a field, then $\mathbf{A}(\mathfrak{F})$ satisfies P_2.*

In this section we have followed the terminology from Blumenthal, 1961. Other names for the First and Third Desargues Properties, used in the literature are respectively the Minor, or Weak, and Major, or Strong, Desargues Properties; likewise for the First and Second Pappus Properties (Szmielew, 1983). Hereafter, by "the Desargues Property" we will mean the Third Desargues Property, and by "the Pappus Property" we will mean the Second Pappus Property. Accordingly, we will speak about *Desarguesian* and *Pappian* affine planes.

Finally, we note that the Desargues and Pappus properties of affine planes have precise analogues for projective planes, satisfying the same relationships with their algebraic counterparts. In fact, the projective versions of Desargues and Pappus properties are simpler, since they need not take into account the cases of parallel vs intersecting lines. For instance, all affine Desargues' properties turn out to be particular cases in projective extensions of affine planes of the *projective Desargues' property* which simply states that "*If two triangles are*

perspective from a point (meaning that the lines determined by the three pairs of corresponding vertices are co-punctual), then they are perspective from a line (meaning that the three intersection points of the pairs of corresponding sides of the two triangles are collinear)." Actually, this property holds in a projective plane iff it can be embedded into a projective 3D-space. Likewise, the two affine Pappus properties are combined in one projective Pappus property. For more details, see e.g. Blumenthal, 1961, Hartshorne, 1967, Mihalek, 1972, Blumenthal and Menger, 1970, Hughes and Piper, 1973.

6.4 Analytic geometry and affine transformations of affine planes over a field

Affine planes with the Pappus Property are close enough to the real affine plane that one can introduce not only coordinatization, but even develop analytic geometry of points and lines in them. In fact, for most of what follows it suffices to assume the Desargues Property, i.e. to consider planes $\mathbf{A}(\mathfrak{F})$, where $\mathfrak{F}$ is a skew field, but to avoid having to deal with the non-commutative multiplication, we assume that $\mathfrak{F}$ is a field. Recall from Sec. 6.3 (see also e.g. Gemignani, 1971, Sec. 3 or Blumenthal, 1961, Sec. V.9) that, given a coordinate system OXY in such an affine plane $\mathbf{A}(\mathfrak{F})$, any line l is determined by an equation $y = ax + m$ if not parallel to the line OY, otherwise by $x = c$, where $a, m, c \in \mathfrak{F}$ are fixed parameters. In either case, the line has a *general equation* $ax + by + c = 0$ where at least one of a, b is not 0, and conversely, every such equation represents a line in $\mathbf{A}(\mathfrak{F})$ in the standard analytic geometric sense. Furthermore, a change of the coordinate system to a new one $O'X'Y'$, with coordinate axes $O'X'$ and $O'Y'$ having equations in the old system respectively $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$, leads to change of the coordinates (x, y) of a given point in the plane according to the following equations:

$$x' = u(a_1x + b_1y + c_1),\ y' = v(a_2x + b_2y + c_2),$$

where $u = (a_1e_x + b_1e_y + c_1)^{-1}$ and $v = (a_2e_x + b_2e_y + c_2)^{-1}$ where (e_x, e_y) are the coordinates of the new unit point I' in the old coordinate system OXY. The non-parallelism of the new coordinate axes is analytically expressed by the condition

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1b_2 - a_2b_1 \neq 0.$$

Thus, change of the coordinate system in $\mathbf{A}(\mathfrak{F})$ can be represented (after multiplying out in the equations above) as a transformation of the plane, determined by *affine equations*:

$$\alpha(x) = ax + by + c,\ \alpha(y) = a'x + b'y + c',$$

where $a, b, c, a', b', c' \in \mathfrak{F}$ are such that $ab' - a'b \neq 0$. Such transformation is called an *affine transformation*, or an *affinity*. Conversely, we will see further that every affine transformation can be viewed as a change of the coordinate system.

It is easy to see that every affine transformation is a collineation on the plane. Moreover, using coordinatization, we can characterize explicitly *all* collineations in an affine plane over a field $\mathbf{A}(\mathfrak{F})$. First, recall that collineations of linear spaces preserve incidence, and therefore parallelism, of lines, and collinearity of points. Actually, a bijection α of the points in the plane is a collineation iff it preserves line parallelism, i.e. if α maps the point P to the point P^α then:

$$AB \parallel CD \;\Leftrightarrow\; A^\alpha B^\alpha \parallel C^\alpha D^\alpha.$$

Now, consider a collineation α in $\mathbf{A}(\mathfrak{F})$ and let OXY be any coordinate system in $\mathbf{A}(\mathfrak{F})$. Since α preserves every line (as a set of points), in particular the unit line, it determines a bijection $h : \mathfrak{F} \longrightarrow \mathfrak{F}$ by sending the point from the unit line (x, x) to the point $(h(x), h(x))$. Recall that addition and multiplication in $\mathfrak{F}$ were geometrically defined in Sec. 6.2 on the unit line by means of the "affine operations" (Sec. 5.7) of taking the intersection point of two lines, producing the line through two points, and producing the line parallel to a given line through a given point. These constructions are preserved by collineations, and therefore h is an *automorphism* of $\mathfrak{F}$. Now, for any point P with coordinates (x, y) in the system OXY, its image under α is the point P^α with coordinates $(h(x), h(y))$ in the system $O'X'Y'$, because the point P can be obtained (see Sec. 6.1) from the points (x, x) and (y, y) by affine operations.

We can now obtain an explicit algebraic characterization of the collineation α. Suppose the images of O, X, Y under α are $O'(c, c'), X'(a+c, a'+c'), Y'(b+c, b'+c')$. (Note that such $a, b, c, a', b', c' \in \mathfrak{F}$ always exist.) It is easy to check that O', X', Y' are non-collinear iff $ab' \neq a'b$. Now, following the construction on Fig.7.2 (or, as a standard exercise in linear algebra) one can compute the coordinates of $\alpha(P)$ in OXY:

$$\alpha(x) = ah(x) + bh(y) + c,\ \alpha(y) = a'h(x) + b'h(y) + c'.$$

Conversely, it is immediate to check that every mapping defined by such equations in a given coordinate system, where h is an automorphism of $\mathfrak{F}$ and $ab' - a'b \neq 0$, is a collineation.

Thus, we have obtained the following (see Gemignani, 1971, Sec. 3):

THEOREM 7.34 *A mapping α in the plane $\mathbf{A}(\mathfrak{F})$, where $\mathfrak{F}$ is a field, is a collineation, iff it can be defined in some coordinate system by equations*

$$\alpha(x) = ah(x) + bh(y) + c,\ \alpha(y) = a'h(x) + b'h(y) + c',$$

where h is an automorphism of $\mathfrak{F}$ and $a, b, c, a', b', c' \in \mathfrak{F}$ are such that $ab' \neq a'b$.

Therefore, every collineation of $\mathbf{A}(\mathfrak{F})$ is uniquely determined by its action on any three non-collinear points O, X, Y in the plane, i.e. by their images O', X', Y', and any mapping in $\mathbf{A}(\mathfrak{F})$ that sends the three non-collinear points O, X, Y respectively to three non-collinear points $O'(c, c'), X'(a + c, a' + c'), Y'(b+c, b'+c')$ can be uniquely extended to a collineation of $\mathbf{A}(\mathfrak{F})$ defined by the equations above.

We now see that affine transformations form a special case of collineations, corresponding to the identity automorphism of $\mathfrak{F}$.

Note that the affinities of a plane form a subgroup of its group of collineations. In the case when $\mathfrak{F}$ is rigid, i.e. has no non-trivial automorphisms, as is the field of reals $\mathbb{R}$, every collineation in $\mathbf{A}(\mathfrak{F})$ is an affinity, but in general this need not be the case, e.g. (see Gemignani, 1971, Sec. 3.2) the complex conjugate mapping $h(z) = \overline{z}$ is an automorphism of the field of complex numbers $\mathbb{C}$, and therefore any collineation of $\mathbf{A}(\mathbb{C})$ associated with h , e.g. $(x, y) \to (\overline{x}, \overline{y})$, is not an affinity.

A particular case of affine transformations is *dilation* (or *dilatation*). This is a collineation δ which sends every line to a parallel one, i.e.,

$$A^\delta B^\delta \parallel AB.$$

The set of dilations of an affine plane $\mathfrak{A}$ will be denoted as $\mathsf{Dil}(\mathfrak{A})$. It can be easily shown (see Gemignani, 1971, Behnke et al., 1974, Coxeter, 1969, Hughes and Piper, 1973) that every dilation of $\mathbf{A}(\mathfrak{F})$ can be defined in a suitable coordinate system by equations

$$\alpha(x) = ax + c,\ \alpha(y) = ay + c',$$

for some $a, c, c' \in \mathfrak{F}$ such that $a \neq 0$. Therefore, if a dilation is different from the identity dilation ι (for which every point is a fixed point), then it has either no fixed points (if $a = 1$ and $(c, c') \neq (0, 0)$) or exactly one fixed point $C((1 - a)^{-1}c, (1 - a)^{-1}c')$; accordingly, it will be called respectively a *translation* or *homothety with center C*. The set of all translations of $\mathfrak{A}$ will be denoted $\mathsf{Tr}(\mathfrak{A})$ while the set of homotheties with center C will be denoted $\mathsf{Ht}_{\mathrm{C}}(\mathfrak{A})$; the set of all homotheties will be denoted $\mathsf{Ht}(\mathfrak{A})$. When the plane is fixed, we will sometimes omit it from these notations. The identity dilation is taken by definition as both a translation as well as a homothety with any point taken as its center.

If α is any non-identity translation and P, Q are any points in the plane then $PP^\alpha \parallel QQ^\alpha$. Hence, all lines PP^α lie in the same "direction", which will be called the *direction* of the translation α. Given any line l, we define the *set of*

translations with direction l:

$$\mathsf{Tr}_l := \{\alpha \in \mathsf{Tr} : \alpha \text{ has direction } l\} \cup \{\iota\}.$$

Another important class of affine transformations is the class of *rotations*, given by equations

$$\alpha(x) = ax + by,\ \alpha(y) = a'x + b'y,$$

for some $a, b, a', b' \in \mathfrak{F}$ such that $\begin{vmatrix} a & b \\ a' & b' \end{vmatrix} = 1$. Rotations are not dilations, but have a fixed point, viz. $O(0,0)$.

For any line l and any point O in a given affine plane, we have:

$$\mathsf{Tr}_l \leq \mathsf{Tr} \leq \mathsf{Dil} \leq \mathsf{Aut},\ \ \mathsf{Ht}_{\mathrm{O}} \leq \mathsf{Dil} \leq \mathsf{Aut},$$

where $\leq$ means subgroup.

A set of dilations D is *transitive* on a set of points S if, for every $A, B \in S$ the equation $A^\delta = B$ has a solution for δ in D. Given any line l, the set Tr_l will be called transitive if it is transitive on the set of points in the line l. The set Tr will be called transitive if it is transitive on the entire universe of points. The set Ht_{O} will be called transitive when, for every line l containing O, the set Ht_{O} is transitive on the set of points $l \backslash \{O\}$. Finally Ht will be called transitive when Ht_{O} is transitive for every point O.

A set of dilations D is called *commutative* when compositions of dilations in D commute. It can be shown that (i) the set Tr will be commutative if and only if the set Tr_l is commutative for every l, and (ii) if Tr is transitive then it is also commutative.

Transitivity and commutativity of the dilations of a given affine plane are closely related to the Desargues and Pappus properties satisfied in that plane. For a more comprehensive discussion on these relations, the reader is referred to Szmielew, 1983, from where we cite the following.

THEOREM 7.35 *An affine plane satisfies*

i) D_1 *iff the set* Tr *is transitive;*

ii) D_3 *iff the set* Ht *is transitive;*

iii) P_2 *iff the set* Ht_{O} *is transitive and commutative for every point* O.

7. On the first-order theories of affine and projective spaces

Here we will discuss some logical results about definability in affine spaces and axiomatization and decidability of the first-order theories of affine and projective spaces, that can be obtained as consequences from the method of coordinatization.

7.1 On affine relations in affine spaces

Two lines x and y in an affine structure are called *crossing* or *skew*, denoted $x \bowtie y$, if they are not incident and not parallel. Thus, each of the relations **Int** , $\parallel$ and $\bowtie$ in affine spaces is definable in terms of incidence between a point and a line. Therefore, affine spaces and planes can be defined with these relations taken as primitives, but that would not enhance the expressiveness of the language.

Note that every relation in an affine plane definable in terms of incidence alone is preserved under collineations. Therefore, using collineations one can show e.g. that orthogonality of lines in $\mathbb{R}^n$ is not definable in terms of the relation of incidence alone, for any $n \geq 2$. Indeed, the mapping in $\mathbb{R}^n$ that halves the first coordinate of a point is clearly a collineation, but it does not preserve orthogonality.

Further, we can define an *affine relation* in affine planes as one which is preserved under affine transformations. Thus, incidence and parallelism are affine relations, while orthogonality is not.

Note on the other hand, that many not obviously affine concepts can be defined in affine terms, or constructed with purely affine means, in affine planes satisfying special additional properties, e.g. in $\mathbf{A}(\mathbb{R})$. For example, the equidistance relation on strictly parallel line segments, denote it here as $\equiv_1$, is given by the formula

$$X_1X_2 \equiv_1 Y_1Y_2 \;\Leftrightarrow\; X_1 = X_2 \wedge Y_1 = Y_2 \vee \Big(\mathbf{l}(X_1, X_2) \parallel \mathbf{l}(Y_1, Y_2)$$
$$\wedge \Big(\mathbf{l}(X_1, Y_1) \parallel \mathbf{l}(X_2, Y_2) \vee \mathbf{l}(X_1, Y_2) \parallel \mathbf{l}(X_2, Y_1)\Big)\Big)$$

(see Fig. 7.13), while equidistance on arbitrary parallel line segments, denote it here as $\equiv_2$, is given by the formula

$$X_1X_2 \equiv_2 Y_1Y_2 \;\Leftrightarrow\; \exists Z_1 \exists Z_2\, (X_1X_2 \equiv_1 Z_1Z_2 \wedge Y_1Y_2 \equiv_1 Z_1Z_2)$$

(see Fig. 7.14). In particular, the midpoint operation between two points, $\oplus$, is given by the formula

$$X = Y_1 \oplus Y_2 \;\Leftrightarrow\; XY_1 \equiv_2 XY_2.$$

As we will note in Sec. 8, betweenness of points in $\mathbb{R}^n$ for $n \geq 1$ is an affine relation, but is not definable in terms of incidence alone, because it may not be preserved by collineations which are not affinities.

7.2 Coordinatization as logical interpretation

Every coordinatization of an affine plane $\mathfrak{A}$ defines an interpretation of $\mathfrak{A}$ in the corresponding ternary ring $\mathfrak{T} = (\Gamma; 0, 1, T)$ attached to it. Indeed, if we treat

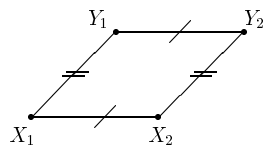


Figure 7.13.

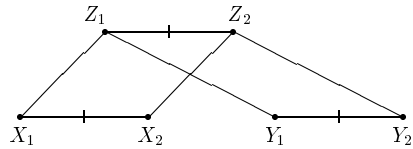


Figure 7.14.

the plane as a collinearity structure $\langle \mathbf{Po}, \mathbf{Col} \rangle$, then $\mathfrak{A}$ can be 2-dimensionally interpreted in $\mathfrak{T}$ as follows:

i) The domain of $\mathfrak{A}$ interpreted in $\mathfrak{T}$ is given by the formula

$$\psi(x_1, x_2) := (x_1 = x_1).$$

ii) The pair (a, b) in Γ^2 is mapped to the point $P(a, b)$ in $\mathfrak{A}$.

iii) Collinearity of points in $\mathfrak{A}$ is given by the formula

$$\begin{aligned}\psi_{\mathbf{Col}}(x_1, y_1, x_2, y_2, x_3, y_3) := {} & \wedge_{i,j=1,2,3}\, x_i = x_j \\ & \vee \exists m \exists c\, (\wedge_{i=1,2,3}\, y_i = T(m, x_i, c))\,.\end{aligned}$$

If the plane is regarded as a two-sorted incidence structure $\langle \mathbf{Po}, \mathbf{Li}, \mathbf{I} \rangle$ then $\mathfrak{A}$ can be 4-dimensionally interpreted in $\mathfrak{T}$ as follows:

i) The domain of $\mathfrak{A}$ interpreted in $\mathfrak{T}$ is given by the formula

$$\psi(x_1, x_2, x_3, x_4) := (x_1 = x_1).$$

ii) The quadruple (a, b, c, d) in Γ^4 is mapped to the point $P(a, b)$ if $(a, b) = (c, d)$, and to the line determined by the points $P(a, b)$ and $Q(c, d)$ otherwise.

iii) Incidence of a point and a line in $\mathfrak{A}$ is given by the formula

$$\begin{aligned}&\psi_{\mathbf{I}}(x_1, x_2, x_3, x_4, y_1, y_2, y_3, y_4) := \\ &\Big((x_1 = x_3 \wedge x_2 = x_4) \wedge (y_1 \neq y_3 \vee y_2 \neq y_4)\Big) \wedge \\ &\quad \Big(\exists m \exists c \big(x_2 = T(m, x_1, c) \wedge y_2 = T(m, y_1, c) \\ &\qquad \wedge\, y_4 = T(m, y_3, c)\big) \vee (x_1 = y_1 \wedge y_1 = y_3)\Big)\end{aligned}$$

(informally: $\bar{x}$ is a point, $\bar{y}$ is a line and $\bar{x}$ lies on $\bar{y}$).

Alternatively, lines in planes satisfying D_1 can be interpreted in the coordinate ring as triples of coefficients, by their general equations.

Conversely, any ternary ring can be interpreted in the affine plane over it, by taking the points on the unit line as a domain of the interpretation, and defining addition and multiplication by means of first-order formulae in the language of incidence, constructed following the geometric description of these operations, as described above and illustrated by Fig. 7.5 and Fig. 7.6.

7.3 Decidability and undecidability of affine and projective theories

The interpretations between affine planes and coordinate rings enable effective translation of first-order formulae from one to the other language and transfer of various logical properties between the first order theories of these classes of structures.

THEOREM 7.36 *For every affine plane $\mathfrak{A}$ and every ternary ring $\mathfrak{T}$:*

1. *If the first order theory of $\mathbf{A}(\mathfrak{T})$ is decidable then the first order theory of $\mathfrak{T}$ is decidable, too.*
2. *If the first order theory of $\mathbf{T}_{OXY}(\mathfrak{A})$ is decidable for some coordinate system OXY in $\mathfrak{A}$, then the first order theory of $\mathfrak{A}$, expanded with point-constants for O, X, Y, is decidable, too.*

Given a class of ternary rings $\mathcal{T}$ we denote by $\mathbf{A}(\mathcal{T})$ the class of affine planes over these rings; likewise, given a class of affine planes $\mathcal{A}$ with the Desargues property, we denote by $\mathbf{T}(\mathcal{A})$ the class of ternary rings attached to the planes in $\mathcal{A}$.

THEOREM 7.37 *For every class of ternary rings $\mathcal{T}$ and every class of affine planes $\mathcal{A}$ satisfying the Desargues property the following holds:*

1. *If the first order theory of $\mathbf{A}(\mathcal{T})$ is decidable then the first order theory of $\mathcal{T}$ is decidable, too.*
2. *If the first order theory of $\mathbf{T}(\mathcal{A})$ is decidable, then the first order theory of $\mathcal{A}$ is decidable, too.*

As shown in Tarski, 1949a, Tarski and Mostowski, 1949, Tarski, 1949b and Tarski et al., 1953, the first order theories of all fields, and the field of rationals, are undecidable, while the first order theory of real closed fields, being the same as the first order theory of the field of reals is complete and decidable. Therefore, we obtain the following.

COROLLARY 7.38

1. *The first order theories of all Pappian affine planes, and of the rational affine plane are undecidable.*
2. *The first order theories of all affine planes over real closed fields, and of the real affine plane are decidable.*

A simple argument shows that if a first-order theory T has an undecidable extension by means of finitely many axioms T', then it is itself undecidable.

Indeed, let ϕ be the conjunction of all axioms extending T to T'. Then for any sentence ψ, $T' \vdash \psi$ iff $T \vdash \phi \rightarrow \psi$, hence any decision method for T yields a decision method for T'. Thus, we obtain the following results:

COROLLARY 7.39 *The following first-order theories are undecidable: the theory of all Desarguesian planes; the theory of all affine planes; the theory of all affine spaces; the theory of all linear spaces; the theory of all incidence structures.*

Analogous results hold for first-order theories of projective planes and spaces (see Ziegler, 1982).

7.4 On the axiomatizations of the first-order theories of the real projective and affine planes

The real affine plane $\mathbf{A}(\mathbb{R})$ is simply the Euclidean plane with the standard points, lines and incidence relation. The real projective plane $\mathbf{P}(\mathbb{R})$ can be obtained from $\mathbf{A}(\mathbb{R})$ by the extension construction described earlier, but also e.g. by the well-known *central projection* of the affine plane onto a sphere touching that plane (see Coxeter, 1969).

Here we briefly discuss the questions: *what are the first-order axiomatizations of the real projective and affine planes* $\mathbf{P}(\mathbb{R})$ *and* $\mathbf{A}(\mathbb{R})$ *in the language with incidence?*

The first-order theory of $\mathbb{R}$ has a well-known axiomatization (the theory of real-closed fields, see e.g. Tarski, 1967, Chang and Keisler, 1973). It extends the axioms for fields with the following axiom schemes:

RealFields : -1 *is not a sum of squares*:

$$\forall x_1 \ldots \forall x_n \neg(x_1^2 + \ldots + x_n^2 = -1)$$

for every integer $n > 0$.

RealClosedFields : *Every polynomial of odd degree has a zero*:

$$\forall a_0 \ldots \forall a_n (\neg a_n = 0 \rightarrow \exists x (a_0 + a_1 x + \ldots + a_n x^n = 0))$$

for every odd integer $n > 0$.

PythagoreanFields :

$$\forall x \exists y (y^2 = x \vee y^2 = -x).$$

In view of the mutual interpretability between $\mathbb{R}$ and each of $\mathbf{P}(\mathbb{R})$ and $\mathbf{A}(\mathbb{R})$, and the uniqueness of the coordinate field for each of these planes, translating the axioms above to the geometric language should in principle

suffice to axiomatize their first-order theories. Still, it is natural to search for *explicit and geometrically meaningful* axiomatizations of the real projective and affine planes, rather than a translation of the axioms of real closed fields to the geometric language.

When betweenness is added to the language, such a complete axiomatization (involving an infinite axiom scheme of continuity) for the real affine plane has been obtained by Szczerba and Tarski, 1965 and Szczerba and Tarski, 1979, and will be presented in Sec. 8. The language with betweenness, however, is substantially more expressive, so the question is: what affine properties of $\mathbb{R}^2$ can be expressed in terms of incidence alone, in projective and affine settings.

We already know that there are rather non-trivial *universal* properties true in $\mathbf{P}(\mathbb{R})$ and $\mathbf{A}(\mathbb{R})$, such as the Pappus property which guarantees that their coordinate ring is a field.

Furthermore, there is a geometrically natural axiom, known as the *Fano* axiom, which is true in $\mathbf{P}(\mathbb{R})$ but does not follow from the Pappus property. In order to state the Fano axiom, we define a *complete quadrangle* in a projective plane to be a configuration of 7 points and 6 lines obtained as follows: take 4 points A, B, C, D, no three of which are collinear, consider the 6 lines determined by pairs of these points, and add the 3 "diagonal points" of intersection $\mathbf{P}(AB, CD)$, $\mathbf{P}(AC, BD)$ and $\mathbf{P}(AD, BC)$.

Fano : The three diagonal points in any complete quadrangle are never collinear.

The Fano axiom is true in every projective plane over a field of characteristic different from 2 (see e.g. Coxeter, 1969), but it fails in the Fano plane. An affine version of the Fano axiom can be formulated, too. In the particular case where the denied collinearity is along the infinite line, it claims precisely that the diagonals of every parallelogram in the plane must intersect.

The Pappus property and Fano axiom are the only additional axioms to those for projective planes offered in Coxeter, 1969, Sec. 14.1 for the real projective plane. However, as shown in an exercise following Coxeter, 1969, Sec. 14.1, for every prime p there is a *finite* projective plane $\mathrm{PG}(2, p)$ of $p^2 + p + 1$ points and as many lines satisfying all these axioms, so this system is far from complete.

In fact, there are infinitely many other geometric axioms which should be added to the theory of Pappian planes, in order to obtain the complete theory of $\mathbf{A}(\mathbb{R})$, because it follows from results in Szczerba and Tarski, 1979 that the latter theory is not finitely axiomatizable. For further discussion and results on this, see von Plato, 1995, Pambuccian, 2001b. Still, the question of finding explicit and geometrically natural axiomatizations for $\mathbf{P}(\mathbb{R})$ and $\mathbf{A}(\mathbb{R})$ apparently remains, as far as we know, unclosed. The same questions can be raised about the first-order theories of the n-dimensional real affine spaces $\mathbf{A}(\mathbb{R}^n)$ (with $n \geq 3$) generated over the field of reals.

8. Betweenness structures and ordered affine planes

We now consider the geometric language in which the only primitive relation is the ternary relation $\mathbf{B}$ of betweenness on points. $\mathbf{B}(XYZ)$ means that the points X, Y and Z are collinear and Y lies between X and Z (with possibly Y coinciding with X or Z). The language consisting of the betweenness relation is significantly more expressive than the language with collinearity, and yet, as we will see later, betweenness is very much an affine notion so that it makes sense to add it to the language of affine geometries, as was done by Tarski.

From the results in Sec. 6.4 is easy to see that collineations on the real affine plane preserve ratios of parallel line segments and consequently also betweenness on points, so that axiomatizing the real affine plane using betweenness does not leave the realm of affine geometry.

8.1 Betweenness structures and ordered geometry

A structure $(S; \mathbf{B})$ consisting of a non-empty set S and a ternary relation $\mathbf{B}$ on S will be called a *linear betweenness structure* provided the following axioms are satisfied:

B1 : $\forall X \forall Y \forall Z (\mathbf{B}(XYZ) \vee \mathbf{B}(YZX) \vee \mathbf{B}(ZXY))$ (connectivity)

B2 : $\forall X \forall Y (\mathbf{B}(XYX) \rightarrow X = Y)$

B3 : $\forall X \forall Y \forall Z (\mathbf{B}(XYZ) \rightarrow \mathbf{B}(ZYX))$ (symmetry)

B4 : $\forall U \forall X \forall Y \forall Z (\mathbf{B}(UXY) \wedge \mathbf{B}(UYZ) \rightarrow \mathbf{B}(XYZ))$

(inner transitivity)

B5 : $\forall U \forall X \forall Y \forall Z (X \neq Y \wedge \mathbf{B}(UXY) \wedge \mathbf{B}(XYZ) \rightarrow \mathbf{B}(UYZ))$

(outer transitivity)

The relation $\mathbf{B}$ is called the *betweenness relation* of the linear betweenness structure. In Szmielew, 1983 it is shown that these axioms are independent, although when the cardinality of the set S is different from 4, the axiom B5 becomes redundant.

Linear orderings and linear betweenness structures are closely related as follows. Let a linear ordering $(S; \leq)$ be given. Then the structure $(S; \mathbf{B}_{\leq})$ with $\mathbf{B}_{\leq}$ defined as

$$\mathbf{B}_{\leq}(XYZ) \;\Leftrightarrow\; X \leq Y \leq Z \vee Z \leq Y \leq X$$

is a linear betweenness structure. Conversely, let a linear betweenness structure $(S; \mathbf{B})$ be given and take any distinct $A, B, C \in S$ such that $\mathbf{B}(ABC)$. Then the structure $(S; \leq_{\mathbf{B}})$ with $\leq_{\mathbf{B}}$ defined as

$$X \leq_{\mathbf{B}} Y \;\Leftrightarrow\; \big(\mathbf{B}(XYB) \wedge \mathbf{B}(XBC)\big) \;\vee$$

$$\big(\mathbf{B}(XBC) \wedge \mathbf{B}(ABY)\big) \vee \big(\mathbf{B}(ABY) \wedge \mathbf{B}(BXY)\big)$$

is a linear ordering. The purpose of the parameters A, B, and C is to fix the direction of $\leq_{\mathbf{B}}$, since clearly every linear betweenness structure gives rise to a *pair* of mutually converse linear orderings. In fact, the parameters A, B, C are inessential in the sense that if $A_i, B_i, C_i \in S$ are distinct with $\mathbf{B}(A_iB_iC_i)$ $(i = 1, 2)$ then the two pairs of linear orderings determined by these two triples of parameters will be identical. Betweenness structures and linear orderings related as above will be called *adjoint*. Thus, every linear betweenness structure has a pair of mutually converse linear orderings adjoint to it, and every linear ordering has a linear betweenness structure adjoint to it.

A linear betweenness structure $(S; \mathbf{B})$ is *dense* if it satisfies the axiom

$$\forall X_1 \forall X_2 \, (X_1 \neq X_2 \rightarrow \exists Y \, (\mathbf{B}(X_1YX_2) \wedge X_1 \neq Y \wedge X_2 \neq Y)) \, .$$

A linear betweenness structure $(S; \mathbf{B})$ is called *Dedekind complete* if it satisfies the second-order axiom

$$\forall \mathcal{P}_1 \forall \mathcal{P}_2 \Big(\exists Y \mathbf{B}(Y\mathcal{P}_1\mathcal{P}_2) \rightarrow \exists Z \mathbf{B}(\mathcal{P}_1 Z \mathcal{P}_2) \Big),$$

stating that if all points of the set $\mathcal{P}_1$ precede all points of the set $\mathcal{P}_2$, i.e. if $\mathbf{B}(YX_1X_2)$ for all $X_1 \in \mathcal{P}_1$ and $X_2 \in \mathcal{P}_2$, then there is a point which separates $\mathcal{P}_1$ and $\mathcal{P}_2$. Accordingly, a linear ordering $(S; \leq)$ is called *Dedekind complete* if the following second-order axiom is satisfied:

$$\forall \mathcal{P}_1 \forall \mathcal{P}_2 \Big(\bigwedge_{i=1,2} \mathcal{P}_i \neq \emptyset \wedge \mathcal{P}_1 \cup \mathcal{P}_2 = S \wedge \mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset \wedge \mathcal{P}_1 \leq \mathcal{P}_2 \\ \rightarrow \exists X \big(\mathcal{P}_1 \leq X \leq \mathcal{P}_2 \big) \Big).$$

This axiom states that if $\mathcal{P}_1$ and $\mathcal{P}_2$ are non-empty disjoint sets that cover the entire set S and if all elements in $\mathcal{P}_1$ are dominated by all elements in $\mathcal{P}_2$, then there is a point that separates $\mathcal{P}_1$ and $\mathcal{P}_2$. Szmielew, 1983 shows that if a linear betweenness structure $(S; \mathbf{B})$ is adjoint to a linear ordering $(S; \leq)$ then $(S; \mathbf{B})$ will be Dedekind complete if and only if $(S; \leq)$ is Dedekind complete.

A linear betweenness structure (respectively a linear ordering) is called *continuous* if it is dense and Dedekind complete.

Now, a betweenness relation is defined on a collinearity structure $\langle \mathbf{Po}, \mathbf{Col} \rangle$ by defining it on every line in that structure. Thus we deal with a geometric structure $\mathfrak{C} = \langle \mathbf{Po}, \mathbf{Col}, \mathbf{B} \rangle$ such that

1. $\mathbf{B}$ is a linear ternary relation on $\mathbf{Po}$, i.e. $\mathbf{B} \subseteq \mathbf{Col}$;

2. $(\mathbf{l}(X, Y); \mathbf{B})$ is a linear betweenness structure for every line $\mathbf{l}(X, Y)$ from $\mathbf{Li}(\mathfrak{C})$.

Point collinearity can be defined in terms of betweenness:

$$\mathbf{Col}(X_1X_2X_3) := \bigvee_{\neq(i,j,k)} \mathbf{B}(X_iX_jX_k).$$

Then, the axioms for betweenness in a collinearity structure are adjusted by adding the axiom B6 and replacing the axiom B1 with the axiom B7:

B6 : $\forall X \forall Y \forall Z(\mathbf{B}(XYZ) \rightarrow \mathbf{Col}(XYZ))$ (linearity)

B7 : $\forall X_1 \forall X_2 \forall X_3(\mathbf{Col}(X_1X_2X_3) \rightarrow \bigvee_{\neq(i,j,k)} \mathbf{B}(X_iX_jX_k))$ (connectivity on lines)

Consequently, betweenness can serve as the only primitive relation in ordered collinearity structures and their axioms can be phrased exclusively in terms of betweenness. A collinearity structure with a betweenness relation imposed on it will be called an *ordered collinearity geometry*. In case the collinearity structure has dimension ≥ 2, and in particular when dealing with the real collinearity plane, it turns out that the axiom B2 becomes redundant.

The betweenness relation has great expressive power; as will be seen in the next section, betweenness was both Veblen's and Tarski's primitive of choice for formalizing affine notions in first-order logic. For example, given points X and Y one can define the following types of line segment: *closed intervals* $[X, Y] := \{Z : \mathbf{B}(XZY)\}$; open intervals $(X, Y) := [X, Y] \setminus \{X, Y\}$; the ray from X away from Y (when $X \neq Y$) $X/Y := \{Z : \mathbf{B}(YXZ)\}$; the line containing X and Y (when $X \neq Y$) $XY := \{Z : \mathbf{B}(ZXY)\} \cup \{Z : \mathbf{B}(XZY)\} \cup \{Z : \mathbf{B}(XYZ)\}$, etc.

8.2 Definability of betweenness and order in affine planes

Note that even if a linear betweenness structure is defined on every line in a collinearity structure, that may not suffice to have a "global" betweenness relation on the entire structure, satisfying the axioms B2 – B7, because the linear betweenness relations may not be synchronizable across the structure. To guarantee that, we should guarantee that betweenness is preserved under parallel projections between lines. This property is formalized by the following three axioms:

Pasch : (Invariance—see Fig. 7.15.)

$$\Big(\neg\mathbf{Col}(X_1X_2X_3Y_1Y_2Y_3) \wedge \mathbf{Col}(X_1X_2X_3) \wedge \mathbf{Col}(Y_1Y_2Y_3)$$
$$\wedge\, \mathbf{B}(X_1X_2X_3) \wedge \textstyle\bigwedge_{i,j=1,2,3} X_iY_i \parallel X_jY_j\Big) \rightarrow \mathbf{B}(Y_1Y_2Y_3)$$

oPasch : (Outer invariance—see Fig. 7.16.)

$$\Big(\neg\mathbf{Col}(X_1X_2X_3Y_2Y_3) \wedge \mathbf{Col}(X_1X_2X_3) \wedge \mathbf{Col}(X_1Y_2Y_3)$$
$$\wedge\, \mathbf{B}(X_1X_2X_3) \wedge X_2Y_2 \parallel X_3Y_3\Big) \;\rightarrow\; \mathbf{B}(X_1Y_2Y_3)$$

iPasch : (Inner invariance—see Fig. 7.17.)

$$\Big(\neg\mathbf{Col}(X_1X_2X_3Y_1Y_3) \wedge \mathbf{Col}(X_1X_2X_3) \wedge \mathbf{Col}(Y_1X_2Y_3)$$
$$\wedge\, \mathbf{B}(X_1X_2X_3) \wedge X_1Y_1 \parallel X_3Y_3\Big) \;\rightarrow\; \mathbf{B}(Y_1X_2Y_3)$$

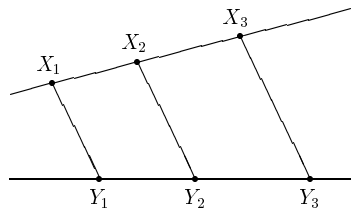


Figure 7.15.

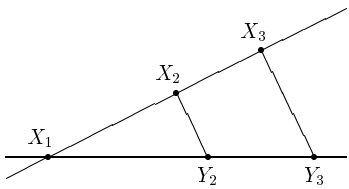


Figure 7.16.

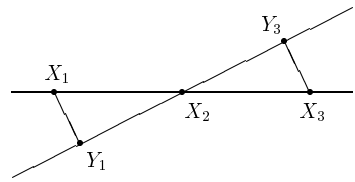


Figure 7.17.

It turns out (see Szmielew, 1983) that in an ordered affine plane, not only are the three Pasch axioms pairwise equivalent, but they are also all equivalent to the Pasch axiom Ax.B5 to be described below.

We now briefly investigate the relationship between ordered affine planes and ordered division rings (see Szmielew, 1983 for details). Let an ordered strong left division ring $\mathfrak{F} = (F; 0, 1, +, \cdot, \leq)$ be given, and let $\mathfrak{F}^- = (F; 0, 1, +, \cdot)$ be the unordered reduct of $\mathfrak{F}$ and $\mathbf{B}$ be the betweenness relation on F adjoint to $\leq$. Then $\mathbf{A}(\mathfrak{F}^-)$ is an affine plane which satisfies the axiom D_1. We can treat $\mathbf{A}(\mathfrak{F}^-)$ as a collinearity structure $\langle F^2, \mathbf{Col}\rangle$. Put $\mathbf{A}(\mathfrak{F}) = \langle F^2, \mathbf{Col}, \mathbf{B}_{\mathfrak{F}}\rangle$, where $\mathbf{B}_{\mathfrak{F}}$ is a ternary relation on the points in $\mathbf{A}(\mathfrak{F}^-)$ defined by stipulating that for any $A = (x_A, y_A), B = (x_B, y_B), C = (x_C, y_C) \in F^2$,

$$\mathbf{B}_{\mathfrak{F}}(ABC) \;\text{ iff }\; \mathbf{Col}(ABC) \;\&\; \mathbf{B}(x_Ax_Bx_C) \;\&\; \mathbf{B}(y_Ay_By_C).$$

On the other hand, let $\mathfrak{A} = \langle \mathbf{Po}, \mathbf{Col}, \mathbf{B} \rangle$ be any ordered affine plane satisfying the axiom D_1 and let $\mathfrak{A}^- = \langle \mathbf{Po}, \mathbf{Col} \rangle$ be the unordered reduct of $\mathfrak{A}$. Fixing some coordinate system OXY in $\mathfrak{A}^-$, the ternary ring $\mathfrak{F}_{OXY}(\mathfrak{A}^-) = (F; 0, 1, +, \cdot)$ attached to $\mathfrak{A}^-$ will be a strong left division ring. Since $\mathbf{B}$ is a betweenness relation in $\mathfrak{A}$ then $\mathbf{B}$ will be a betweenness relation on every line in $\mathfrak{A}$, and hence also on the set F. Thus $(F; \mathbf{B})$ is a linear betweenness structure and it will have a pair of mutually converse linear orderings adjoint to it. Suppose $\leq$ is the linear ordering on F adjoint to $\mathbf{B}$ *and such that* $0 \leq 1$. Put $\mathfrak{F}_{OXY}(\mathfrak{A}) = (F; 0, 1, +, \cdot, \leq)$.

The property that every line in a plane contains at least three points can be axiomatized as follows:

$\mathsf{B8}: \ \forall X \forall Y \exists Z (Z \neq X \wedge Z \neq Y \wedge \mathbf{Col}(XYZ))$

Szmielew, 1983 gives the following representation results.

THEOREM 7.40 *Let $\mathfrak{A}$ be an ordered affine plane satisfying* Pasch *and* B8. *If $\mathfrak{A}$ also satisfies D_1 (respectively, D_3; P_2) then $\mathfrak{A} \cong \mathbf{A}(\mathfrak{F})$ for some ordered strong left division ring (respectively, skew field; field) $\mathfrak{F}$.*

THEOREM 7.41 *If $\mathfrak{F}$ is an ordered strong left division ring (respectively, skew field; field) then $\mathbf{A}(\mathfrak{F})$ is an ordered affine plane satisfying the axioms* Pasch, B8, *and D_1 (respectively, D_3; P_2).*

8.3 Axiomatizing betweenness in $\mathbb{R}^2$

Szczerba and Tarski, 1965 and Szczerba and Tarski, 1979 study the affine fragment AE_2, called the *elementary affine Euclidean geometry*, of the Euclidean plane. AE_2 is the elementary geometry formalized in the language with only the betweenness relation $\mathbf{B}$, where a sentence is valid in AE_2 if and only if it is valid in the Euclidean plane E_2. They give a complete axiomatization of AE_2 which will be outlined below (all axioms below are implicitly universally quantified over all occurring free variables).

Ax.B1 : IDENTITY AXIOM

$$\mathbf{B}(XYX) \rightarrow X = Y$$

Ax.B2 : TRANSITIVITY AXIOM

$$Y \neq Z \wedge \mathbf{B}(XYZ) \wedge \mathbf{B}(YZW) \rightarrow \mathbf{B}(XYW)$$

Ax.B3 : CONNECTIVITY AXIOM

$$V \neq W \wedge \mathbf{B}(VWX) \wedge \mathbf{B}(VWY) \rightarrow (\mathbf{B}(VXY) \vee \mathbf{B}(VYX))$$

Ax.B4 : EXTENSION AXIOM

$$\exists X(X \neq Y \wedge \mathbf{B}(XYZ))$$

Ax.B5 : (Outer form of) Pasch axiom

$$\mathbf{B}(XY'Z) \wedge \mathbf{B}(YZ'Y') \rightarrow \exists X'(\mathbf{B}(ZX'Y) \wedge \mathbf{B}(XZ'X'))$$

(given a triangle $YY'Z$, a point X on the extension of the side $Y'Z$ and a point Z' on the inner side (with respect to X) of the triangle, the line XZ' must intersect the triangle in its outer side (with respect to X) $|YZ|$—see Fig. 7.18.)

Ax.B6 : Desargues axiom

$$\begin{aligned}\neg\mathbf{Col}(TXY) \wedge \neg\mathbf{Col}(TXZ) \wedge \neg\mathbf{Col}(TYZ) \wedge \mathbf{B}(TXX') \wedge \mathbf{B}(TYY') \\ \wedge \mathbf{B}(TZZ') \wedge \mathbf{B}(YXU) \wedge \mathbf{B}(Y'X'U) \wedge \mathbf{B}(XZW) \wedge \mathbf{B}(X'Z'W) \\ \wedge \mathbf{B}(YZV) \wedge \mathbf{B}(Y'Z'V) \ \rightarrow \ \mathbf{B}(UVW)\end{aligned}$$

(triangles perspective from a point are perspective from a line—see Fig. 7.19.)

Ax.B7 : Lower 2-dimensional axiom

$$\exists X \exists Y \exists Z(\neg\mathbf{B}(XYZ) \wedge \neg\mathbf{B}(YZX) \wedge \neg\mathbf{B}(ZXY))$$

Ax.B8 : Upper 2-dimensional axiom (See Fig. 7.20.)

$$\begin{aligned}\exists V((\mathbf{B}(YVZ) \wedge \mathbf{Col}(XVW)) \vee (\mathbf{B}(XVZ) \wedge \mathbf{Col}(YVW)) \\ \vee (\mathbf{B}(XVY) \wedge \mathbf{Col}(ZVW)))\end{aligned}$$

As.B9 : Elementary continuity axiom schema

$$\begin{aligned}\forall \overline{W}(\exists U \forall X \forall Y(\varphi(X, \overline{W}) \wedge \psi(Y, \overline{W}) \rightarrow \mathbf{B}(UXY)) \ \rightarrow \\ \exists V \forall X \forall Y(\varphi(X, \overline{W}) \wedge \psi(Y, \overline{W}) \rightarrow \mathbf{B}(XVY)))\end{aligned}$$

The variables $\overline{W}$ are distinct from U, V, X, Y, and $\varphi(X, \overline{W})$ and $\psi(Y, \overline{W})$ are first-order formulae over $\mathbf{B}$ with free variables only amongst $X, \overline{W}$ in the case of φ, and $Y, \overline{W}$ in the case of ψ. This schema comprises the *parametrically first-order definable instances of the full second order continuity axiom* (see Ax.11 further). Note that As.B9 is an infinite schema, and it cannot be replaced by a finite one, as Tarski has shown.

The axiom system given so far is denoted GA_2 and the geometry it describes is called by Szczerba and Tarski the *general affine geometry*. It does not reflect Euclid's parallel postulate at all and it is shown in Szczerba and Tarski, 1979 that:

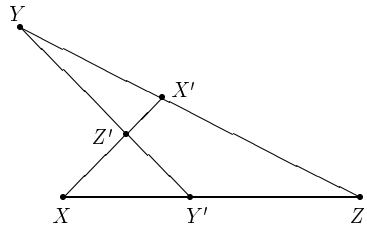


Figure 7.18.

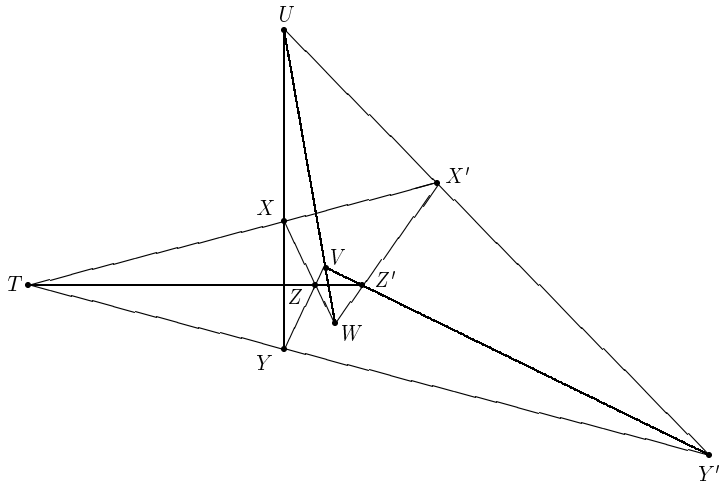


Figure 7.19.

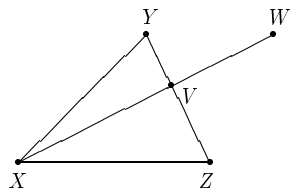


Figure 7.20.

- GA_2 is incomplete and has continuum many complete extensions.
- In particular, GA_2 is a proper subtheory of the *elementary affine absolute geometry* AA_2 (the affine fragment in the language of $\mathbf{B}$ of the absolute geometry A_2, which is the reduct of the Euclidean geometry obtained by dropping Euclid's parallel postulate).

- However, GA_2 is complete with respect to universal sentences, i.e. if a universal sentence σ is true in *some* model of GA_2 then σ is true in *every* model of GA_2.
- GA_2 is not finitely axiomatizable.
- GA_2 is hereditarily undecidable, meaning that both GA_2, as well as all its subtheories, are undecidable.
- GA_2 is decidable with respect to inductive sentences. Therefore, GA_2 is not an inductive theory.

Here is a form of Euclid's postulate in the language of $\mathbf{B}$:

Ax.E : EUCLID'S AXIOM

$$Z \neq V \wedge \mathbf{B}(ZVT) \wedge \mathbf{B}(UVW) \rightarrow$$
$$\exists X \exists Y (\mathbf{B}(ZUX) \wedge \mathbf{B}(ZWY) \wedge \mathbf{B}(YTX)).$$

The axiom Ax.E says that through any point T in the interior of an angle there is a line intersecting both sides of that angle (see Fig. 7.21).

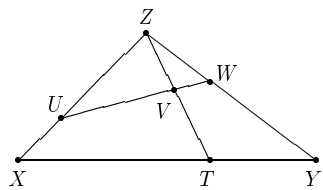


Figure 7.21.

Szczerba and Tarski show that adding that axiom Ax.E to GA_2 renders a complete axiomatization of the elementary affine Euclidean geometry AE_2, and hence also a complete axiomatization of the real affine plane.

Finally, here is a representation result for models of GA_2. Let $\mathfrak{F} = (F; 0, 1, +, \cdot, \leq)$ be any ordered field and define $\oplus$ and $\odot$ as

$$(x_1, y_1) \oplus (x_2, y_2) := (x_1 + x_2, y_1 + y_2), \quad (x_1, y_1) \odot \alpha := (x_1 \cdot \alpha, y_1 \cdot \alpha),$$

for any $(x_1, y_1), (x_2, y_2) \in F^2$ and $\alpha \in F$. We can define a betweenness relation $\mathbf{B}_{\mathfrak{F}}$ in the Cartesian square F^2 over the field $\mathfrak{F}$ as follows: given $a, b, c \in F^2$, we stipulate that

$$\mathbf{B}_{\mathfrak{F}}(abc) \Leftrightarrow b = [a \odot (1 - \lambda)] \oplus [c \odot \lambda]$$

for some $\lambda \in F$ with $0 \leq \lambda \leq 1$. The structure $\mathbf{A}(\mathfrak{F}) = (F^2; \mathbf{B}_{\mathfrak{F}})$ thus formed will be called the *affine plane over the ordered field* $\mathfrak{F}$. Using the

class of all interiors of triangles as a basis, we define a topology on the set F^2. Now let S be any non-empty, convex, open subset of F^2. The structure $\mathbf{A}(\mathfrak{F}; S) = (F^2|_S; \mathbf{B}_{\mathfrak{F}}|_S)$ will be called the *S-restricted affine plane over $\mathfrak{F}$*. A plane over some field will simply be called a *restricted affine plane* if it is an S-restricted affine plane for some S.

THEOREM 7.42 *1. Every model of GA_2 is isomorphic to a restricted affine plane over some ordered real closed field.*

2. Every restricted affine plane over the ordered field of reals is a model of GA_2.

3. If $\mathfrak{F}$ is an ordered real closed field not isomorphic to the field of reals, then there is a restricted affine plane over $\mathfrak{F}$ which is not a model of GA_2.

9. Rich languages and structures for elementary geometry

We will call a geometric language *rich* if the whole of elementary geometry in $\mathbb{R}^n$ is definable in that language. Perhaps the first study on rich primitive notions in elementary geometry is Pieri, 1908, where the Pieri relation Δ is introduced, defined as

$$\Delta(XYZ) := \|XY\| = \|XZ\|\,,$$

meaning that the configuration of points XYZ forms an isosceles triangle with base $|YZ|$ (in the degenerate cases either $Y = Z$ or X is the midpoint of $|YZ|$). Pieri showed that Δ can be used as the *only* primitive relation in $\mathbb{R}^n$ for $n \geq 2$. This result easily implies the richness of many other relations in terms of which Δ is definable, for example the ternary relation of *closer-than*

$$|XY| \leq |XZ|,$$

which states that either the point Y is closer to X than what Z is to X or that Y and Z lie equally far from X. This furthermore implies that the quaternary relation *shorter-than*

$$|XY| \leq |ZU|,$$

which states that the line segment $|XY|$ is shorther than the segment $|ZU|$, or that they are of equal length, is also rich.

Veblen, 1904 considered the two primitive relations of *betweenness* $\mathbf{B}$ and *equidistance* $\equiv$ (or δ), which are the same primitives that Tarski later used. Veblen showed that these primitives are sufficient for the elementary geometry, although he believed to have proved, falsely, that the relation of equidistance is definable in terms of the relation of betweenness (see Tarski and Givant, 1999). In fact, using the coordinatization of the Euclidean plane, and applying Padoa's method, it is easy to see that the equidistance relation $\equiv$ is *not* definable in terms

of betweenness, not only in first-order languages, but even in higher-order logic. Indeed, the linear transformation $f : \mathbb{R}^2 \to \mathbb{R}^2$, defined by $f(x, y) = (x, 2y)$, preserves betweenness, but not equidistance.

On the other hand, Pieri showed that $\mathbf{B}$ is first-order definable in terms of the quaternary closer-than relation $\leq$ defined above:

$$\mathbf{B}(XYZ) \Leftrightarrow \forall U\, ((|XU| \leq |XY| \wedge |ZU| \leq |ZY|) \to U = Y)\,,$$

meaning that if U and Y are intersection points of spheres with centers X and Z then they must coincide (see Fig. 7.22).

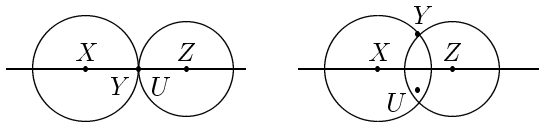


Figure 7.22.

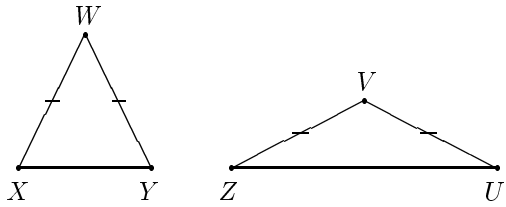


Figure 7.23.

Furthermore, the quaternary shorter-than relation $\leq$ is definable in terms of $\equiv$ as follows:

$$|XY| \leq |ZU| \Leftrightarrow \forall V\, (ZV \equiv UV \to \exists W(XW \equiv YW \wedge YW \equiv UV))\,,$$

meaning that if there is an isosceles triangle with base $|ZU|$ and a given side, then an isosceles triangle with the same side and base $|XY|$ exists too (see Fig. 7.23). It follows that $\equiv$ can be taken as the only primitive for the elementary geometry.

Regarding primitive relations with smaller arities, for every $n \geq 2$, no binary relation can be rich for $\mathbb{R}^n$ (see Beth and Tarski, 1956). Also, as noted earlier, the relation $\mathbf{B}$ alone is not sufficient for the whole of elementary geometry in $\mathbb{R}^2$. Still, Beth and Tarski show in Beth and Tarski, 1956 that the ternary relation $\mathbf{E}$, where $\mathbf{E}(XYZ)$ means that the configuration XYZ forms an equilateral triangle (or degeneratively that the points X, Y and Z coincide), is rich for every $\mathbb{R}^n$ with $n \geq 3$, by expressing Pieri's relation in terms of $\mathbf{E}$ in the system $\mathbb{R}^3$. However, they also show that the relation $\mathbf{E}$ alone is not sufficient for

the geometries $\mathbb{R}^n$ with $n = 1$ or $n = 2$. Similar results hold for the ternary relation specifying that three points determine a rectangular isosceles triangle, and the quaternary relation specifying that four points are the vertices of a square. However, it was shown in (see Beth and Tarski, 1956, Scott, 1956) that the completely symmetric ternary relation $\mathbf{R}$, where $\mathbf{R}(XYZ)$ means that the points X, Y and Z are distinct and form, in some order, a rectangular triangle, is rich for every $\mathbb{R}^n$ with $n \geq 2$.

Schwabhäuser and Szczerba, 1975 investigate *line* relations which can be taken as primitives for the elementary Euclidean geometry. They establish simple rich systems of such relations for every $\mathbb{R}^n$, $n \geq 2$. For the dimension-free Euclidean geometry they show that the binary relation $\dot{\perp}$ of perpendicularity together with the ternary relation $\mathbf{Cop}$ of co-punctuality suffice. For dimensions higher than 3 perpendicularity alone suffices, while for $\mathbb{R}^2$ it does not. For $\mathbb{R}^3$ perpendicularity and the binary relation $\mathbf{Cop}$ of co-punctuality suffice. Later, Kramer, 1993 proves that $\dot{\perp}$ alone is not sufficient as a primitive for $\mathbb{R}^3$, because co-punctuality is not definable there in terms of it. Finally, the question of primitive geometric relations and definability in the case of $\mathbb{R}^1$ turns out to be rather more complicated; the reader is referred to Tarski and Givant, 1999.

9.1 Tarski's system of elementary geometry based on B and δ

In the mid 20th century Tarski developed systematically an axiomatic system for the elementary geometry based on the only primitive concept of *point*, and the two primitive relations *betweenness* $\mathbf{B}$ and *equidistance* $\equiv$ (or δ). Over many years, Tarski and his students refined, simplified, and minimized that system, and a detailed account of that development can be found in Tarski and Givant, 1999, which we follow here for the choice of axioms and notation. Again, all axioms are implicitly universally quantified over all occurring free variables.

Ax.1 : Reflexivity of equidistance.

$$X_1X_2 \equiv X_2X_1$$

Ax.2 : Transitivity of equidistance.

$$(X_1X_2 \equiv Y_1Y_2 \wedge X_1X_2 \equiv Z_1Z_2) \rightarrow Y_1Y_2 \equiv Z_1Z_2$$

Ax.3 : Identity of equidistance.

$$XY \equiv ZZ \rightarrow X = Y$$

Ax.4 : Equal segments construction. (See Fig. 7.24) There is a segment of length $\|Y_1Y_2\|$ beginning at X_1 in the direction of $\overrightarrow{ZX_1}$:

$\exists X_2(\mathbf{B}(ZX_1X_2) \wedge X_1X_2 \equiv Y_1Y_2)$

Ax.5 : FIVE-SEGMENT AXIOM. (See Fig. 7.25) The corresponding line segments built on two congruent triangles are equal:

$$(X \neq Y \ \wedge \ \mathbf{B}(XYZ) \ \wedge \ \mathbf{B}(X'Y'Z') \ \wedge \ XY \equiv X'Y' \ \wedge \ YZ \equiv Y'Z' \\ \wedge \ XW \equiv X'W' \ \wedge \ YW \equiv Y'W') \ \rightarrow \ ZW \equiv Z'W'$$

Ax.7$_1$: (OUTER FORM OF) PASCH AXIOM. (See Ax.B5 above.)

$\mathbf{B}(XY'Z) \wedge \mathbf{B}(YZ'Y') \rightarrow \exists X'(\mathbf{B}(ZX'Y) \wedge \mathbf{B}(XZ'X'))$

Ax.8$^{(2)}$: LOWER 2-DIMENSIONAL AXIOM

$\exists X \exists Y \exists Z(\neg\mathbf{B}(XYZ) \wedge \neg\mathbf{B}(YZX) \wedge \neg\mathbf{B}(ZXY))$

Ax.8$^{(n)}$: LOWER n-DIMENSIONAL AXIOM FOR $n \geq 3$

$$\exists U \exists V \exists W \, \exists X_1 \ldots \exists X_{n-1} \Big(\mathrm{Diff}_{n-1}(X_1 \ldots X_{n-1}) \\ \wedge \neg\mathbf{B}(UVW) \wedge \neg\mathbf{B}(VWU) \wedge \neg\mathbf{B}(WUV) \\ \wedge \bigwedge_{i=2}^{n-1} UX_1 \equiv UX_i \ \wedge \ \bigwedge_{i=2}^{n-1} VX_1 \equiv VX_i \ \wedge \ \bigwedge_{i=2}^{n-1} WX_1 \equiv WX_i \Big)$$

The axiom Ax.8$^{(n)}$ claims that there exist $n-1$ distinct points $X_1, \ldots,$ X_{n-1}, and three non-collinear points U, V, W, each of them equidistant from $X_1, \ldots, X_{n-1}$, which implies that the dimension of the space is at least n. Using these axioms, one can express that the dimension of the space is n:

$$\mathsf{Dim}_n := (\mathsf{Ax.8}^{(n)}) \wedge \neg(\mathsf{Ax.8}^{(n+1)}).$$

Ax.10$_1$: (FORM OF) EUCLID'S AXIOM. (See the axiom Ax.E above.)

$$Z \neq V \wedge \mathbf{B}(ZVT) \wedge \mathbf{B}(UVW) \ \rightarrow \\ \exists X \exists Y(\mathbf{B}(ZUX) \wedge \mathbf{B}(ZWY) \wedge \mathbf{B}(YTX))$$

Ax.11 : SECOND-ORDER CONTINUITY AXIOM. (See also Sec. 8 above.)

$\exists Y(\mathbf{B}(Y\mathbf{X}_1\mathbf{X}_2)) \rightarrow \exists Z(\mathbf{B}(\mathbf{X}_1Z\mathbf{X}_2))$

This axiom says that if all elements of the set $\mathbf{X}_1$ precede all elements of the set $\mathbf{X}_2$ on a line, then there is a point *on* that line which separates $\mathbf{X}_1$ and $\mathbf{X}_2$. This property is not definable in the first-order language of $\mathbf{B}$ and $\equiv$. Its first-order approximation is the corresponding *axiom schema*.

As.11 : CONTINUITY AXIOM SCHEMA. See As.B9 above.

Ax.15 : INNER TRANSITIVITY AXIOM FOR BETWEENNESS.

$$\mathbf{B}(XYW) \wedge \mathbf{B}(YZW) \rightarrow \mathbf{B}(XYZ)$$

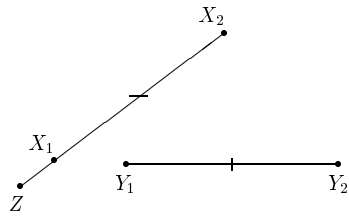


Figure 7.24.

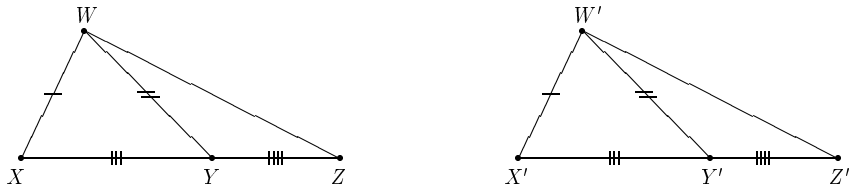


Figure 7.25.

THEOREM 7.43 (TARSKI, 1967) *The set of axioms* Ax.1 – Ax.5, Ax.7_1, Ax.10_1, Ax.15 *plus* Dim_n, *taken with the second order axiom* Ax.11, *characterizes up to isomorphism the full n-dimensional Euclidean geometry $FG^{(n)}$ for every $n \geq 2$. Furthermore, if* As.11 *is taken instead of* Ax.11, *the resulting first-order axiomatic system is a complete axiomatization of the first-order theory of the full n-dimensional Euclidean elementary geometry $EG^{(n)}$ for any $n \geq 2$. If, moreover, the dimension axiom* Dim_n *is omitted, then, according to Scott, 1959, a complete axiomatization of the dimension-free Euclidean geometry is obtained.*

9.2 Decision methods and automated reasoning for elementary geometry

The algebraic approach in geometry goes back at least to Descartes, who introduced the coordinate method in his study of geometry. The first modern development of general algebraic methods used in constructive solutions to classes of geometric problems in affine geometry is due to Hilbert, 1950. However, the first explicit decision method for elementary Euclidean geometry, i.e. a general method for deciding the truth of any first-order sentence in this geometry, was developed in Tarski, 1951. Tarski's decision method is based on a

decision procedure for the first-order theory of the field of real numbers, which is also the first-order theory of the class of real-closed fields. For that theory Tarski established *quantifier elimination*, i.e. it was proved that every first-order sentence formulated in the class of real-closed fields is equivalent, over the class of real-closed fields, to a boolean combination of algebraic equations and inequalities. Equations are simply conjunctions of inequalities, and every inequality can be expressed in the form $t \geq 0$ for some term t in the first-order language of rings (i.e. t will be a polynomial). Therefore, eventually, every first-order sentence of that language is equivalent, over the class of real-closed fields, to a boolean combination of formulae of the type $\exists x(x^2 = t)$, where t is a term not containing x. Subsets of $\mathbb{R}^n$ definable by such formulae are called *semi-algebraic sets*. In particular, Tarski's result implies that the parametrically first-order definable relations in $\mathbb{R}^n$ are precisely the semi-algebraic sets of $\mathbb{R}^n$. For a sketch of an algebraic proof of this result based on Sturm's theorem, see e.g. Hodges, 1993.

Tarski's decision procedure is practically inefficient as it has non-elementary complexity. More efficient, elementary decision procedures were developed later, first by Monk, followed by Solovay, Collins, and others.

Currently there are several well-developed and applied automated theorem proving decision methods for the first-order theory of the field of reals and the theory of elementary geometry.

Probably the most popular decision method for the theory of real closed fields, and the first one amenable to practical automation (it has in fact been implemented), is Collins' method of *Cylindrical Algebraic Decompositions* (CAD), based on quantifier elimination (see Caviness and Johnson, 1998, Collins, 1975, Collins, 1998). Given a boolean combination $\mathcal{B}$ of algebraic equations and inequalities, this algorithm computes a so-called *cylindrical algebraic decomposition* of the solution set of $\mathcal{B}$. This cylindrical algebraic decomposition partitions the solution set of $\mathcal{B}$ into a finite disjoint union of spatial regions called *cells*, which have the property that all polynomials occurring in $\mathcal{B}$ preserve sign on each of these cells. Cells in $\mathbb{R}^n$ can be defined inductively on n as follows: a cell in $\mathbb{R}$ is an open interval or a singleton; a cell in $\mathbb{R}^{n+1}$ is either the graph of a continuous function defined over a cell in $\mathbb{R}^n$, or a region bounded between the graphs $f(\mathbf{x})$ and $g(\mathbf{x})$ of two such functions f and g defined on the same cell in $\mathbb{R}^n$, and such that $f < g$ on that cell. In particular, each of these functions can be taken as $-\infty$ or $+\infty$.

Collins' algorithm has a double exponential worst-case time complexity as a function of the number of variables in $\mathcal{B}$. Later, Heintz et al., 1990 and Renegar, 1992 constructed algorithms for quantifier elimination that are double exponential only in the number of quantifier alternations; see also Davenport and Heintz, 1988. Currently the most efficient algorithms for quantifier elimination

known to us can be found in Basu et al., 1996 and Basu, 1999; the latter employs *uniform quantifier elimination*.

The *Characteristic Set method*, rooted in work by Ritt and later developed independently by Wu (see Chou, 1984, Chou, 1988, Chou, 1990, Chou and Gao, 1990, Wu, 1984, Wu, 1986), and the *Gröbner Basis method*, developed by Buchberger (see Buchberger, 1985, Buchberger et al., 1988, Chou, 1990), work only on problems that can be formalized by systems of equations, and are only complete for algebraically closed fields. A related alternative method, based on Hilbert's *Nullstellensatz*, has been proposed by Kapur, 1986.

Another method, based on ideas coming from quantifier elimination in linear and quadratic formulae over the reals, has been proposed in Dolzmann et al., 1998. Unlike the Characteristic Set and Gröbner Basis methods, it is also applicable to geometric problems in the Euclidean plane and Euclidean n-space whose complex analogues may fail.

Chou, 1984 shows how the Wu-Ritt method of characteristic sets can be applied to finding locus equations, and in Chou, 1987 it is also shown how this method can be used for the mechanical derivation of formulae in elementary geometry. All of these methods require large computational resources and can easily become unfeasible for more complex formulae. For an overview of automated reasoning in geometry see Chou and Gao, 1990.

This concludes our discussion of first-order theories of geometry.

10. Modal logic and spatial logic

In the remainder of this chapter we will survey some modal logics related to classical geometric structures, and from now on we assume some familiarity with basic modal logic and Kripke semantics.

Modal logics related to spatial structures are also considered in Ch. 5 and Ch. 9. The former is mainly oriented towards the topological interpretation of modal logic, whereas the latter deals with the combination of spatial logics and temporal logics.

Basic modal logic. In order to fix the notations and terminology in basic modal logic we will give a short list of definitions and facts. For all notions mentioned without definitions the reader is invited to consult Blackburn et al., 2001 or Hughes and Cresswell, 1996.

Let $\mathcal{C}$ be a class of relational structures of the form $\mathcal{F} = (W, R_1, \ldots, R_n)$, where W is a nonempty set whose elements are usually called *possible worlds*, and $R_1, \ldots, R_n$ are binary relations on W called *accessibility relations*. In modal logic such relational structures are called Kripke frames. We associate with $\mathcal{C}$ a modal language L which is an extension of the standard language for propositional logic with unary connectives $[R_i]$, $i = 1, \ldots, n$, called modal *box*

operators, with the standard definition of a formula, given by the rule:

$$A ::= p \mid \neg A \mid (A \vee B) \mid [R_1]A \mid \ldots \mid [R_n]A.$$

We use the classical abbreviations for "true" ($\top$), "false" ($\bot$), conjunction ($\wedge$), implication ($\rightarrow$), and equivalence ($\leftrightarrow$). We also use the dual, *diamond* operators $\langle R_i \rangle$, defined by $\langle R_i \rangle A = \neg[R_i]\neg A$. Standard modal logic has only one box-modality $\Box$ called "necessity", and the corresponding diamond modality $\Diamond$ is called "possibility".

The semantics of L in a given frame $\mathcal{F} = (W, R_1, \ldots, R_n)$ is based on the notion of *valuation* on $\mathcal{F}$, which is a function V assigning to each proposition letter p a subset $V(p)$ of W. Intuitively, we think of $V(p)$ as a set of possible worlds in which p is true. A pair $\mathcal{M} = (\mathcal{F}, V)$ where V is a valuation on $\mathcal{F}$ is called a *Kripke model* based on $\mathcal{F}$. We define the satisfiability relation $\mathcal{M}, w \models A$ — in words, *the formula A is satisfied at the possible world w of the model $\mathcal{M}$* — as in Ch. 5. In particular we have:

$\mathcal{M}, w \models [R_i]A$ iff $\mathcal{M}, w' \models A$ for all $w' \in W$ such that wR_iw'.

A formula A is true in Kripke model $\mathcal{M}$ iff $\mathcal{M}, w \models A$ for all possible worlds w in $\mathcal{M}$. We say that A is valid in a Kripke frame $\mathcal{F}$ iff A is true in all models defined over $\mathcal{F}$; A is valid in a class $\mathcal{C}$ of Kripke frames iff A is valid in all Kripke frames of $\mathcal{C}$. The set $\mathcal{L}(\mathcal{C})$ of all formulas which are valid in $\mathcal{C}$ is called the logic of $\mathcal{C}$ and the formulas from $\mathcal{L}(\mathcal{C})$ are called the modal laws of $\mathcal{C}$. If $A \in \mathcal{L}(\mathcal{C})$ then we write $\models_{\mathcal{L}(\mathcal{C})} A$.

The above is a semantic definition of a modal logic related to a class $\mathcal{C}$ of frames. Note that one and the same modal logic may be the logic of different classes of frames.

A class $\mathcal{C}$ of frames of the form $(W, R_i, \ldots, R_n)$ is *modally definable* if there exists a formula A such that for every frame $\mathcal{F}$ of the form $(W, R_i, \ldots, R_n)$, $\mathcal{F}$ is in $\mathcal{C}$ iff A is valid in $\mathcal{F}$.

If this is true then we also say that $\mathcal{C}$ is modally definable by A. If the class $\mathcal{C}$ is definable by a first-order condition φ on the relations R_i then we also say that φ is modally definable by A. For instance, reflexivity of a relation R is definable by $[R]p \rightarrow p$, symmetry of R is definable by $p \rightarrow [R]\langle R \rangle p$, and transitivity of R is definable by $[R]p \rightarrow [R][R]p$, where p is a propositional variable. So modal definability is in some sense a way to talk about properties of Kripke frames by means of a propositional language. Let us note that not all first-order properties of Kripke frames are modally definable and that not all modal formulas define a first-order property.

Axiomatically, a modal logic is defined as the smallest set of formulas containing a given set of axioms and closed with respect to a given set of inference rules. The elements of a modal logic $\mathcal{L}$ are called *theorems* of $\mathcal{L}$. If A is a theorem of $\mathcal{L}$ then we write $\vdash_{\mathcal{L}} A$. For instance, the modal logic K_n of the class

of all Kripke frames of the form $(W, R_1, \ldots, R_n)$ has the following axiomatic definition:

Axioms: all substitution instances of classical tautologies, and all formulas of the form $[R_i](A \to B) \to ([R_i]A \to [R_i]B)$ for $i = 1, \ldots, n$.

Inference rules of K_n: Modus ponens "given A and $A \to B$, derive B" and generalization "given A, derive $[R_i]A$".

Axiomatic definitions of other logics can be obtained by adding to the above axiomatic system additional axioms and possibly additional rules of inference. If the axiomatic system does not contain additional inference rules it is called normal. For instance, the logic S4 is an extension of the logic $K_\Box$ with the axiom schemes $\Box A \to A$ (defining reflexivity of R) and $\Box A \to \Box\Box A$ (defining the transitivity of R). The logic S5 is an extension of S4 with the axiom scheme $A \to \Box\Diamond A$ (defining the symmetry of R). The statement of the equivalence of a given semantic definition of a modal logic with a given axiomatic definition is called a *completeness theorem* with respect to the corresponding class of frames. There are different methods for proving completeness theorems. One of them is the so-called *method of canonical models*. An important theorem related to this method is the famous *Sahlqvist theorem* saying that if the axioms of the logic are of a given specified form, then these axioms define first-order conditions, and the logic is canonically complete in the class of frames satisfying these conditions. For instance, the axioms of the logics S4 and S5 are of Sahlqvist's type and hence are "canonical", so S4 is complete in the class of all pre-orders and S5 is complete in the class of all equivalence relations. Furthermore, it is known also that S5 is also complete for the class of all frames with $R = W \times W$, the universal relation in W. Another method for proving completeness theorems is based on the notion of *bounded morphisms*. By means of this method one can prove, for instance, that two different classes of frames, C_1 and C_2, define equal logics $\mathcal{L}(C_1) = \mathcal{L}(C_2)$. If, by some method (for instance by the method of canonical models), one can give a complete axiomatization of $\mathcal{L}(C_1)$, then one automatically obtains a completeness theorem with respect to the class C_2. For the method of canonical models, bounded morphisms, Sahlqvist's theorem and some other methods see Blackburn et al., 2001.

The above-described modal logics contain only unary modal operations. There are also modal logics with binary and in general with n-ary modal operations, called polyadic modalities, with Kripke semantics using relations with arbitrary finite arity. If, for instance, $A \circ B$ is a binary modality then it can be interpreted in frames with a ternary relation $R \subseteq W^3$ as follows:

$\mathcal{M}, w \models A \circ B$ iff there exist w' and w'' such that $wRw'w''$, $\mathcal{M}, w' \models A$ and $\mathcal{M}, w'' \models B$.

Modal logic and applied modal logic. A major aim of modal logic is to study modal logics of different classes of frames, mainly with respect to *modal definability*, *axiomatization*, *decidability* and *complexity*.

The broad applicability of modal logics rests, inter alia, on the fact that, while they are based on propositional languages, every modal formula corresponds in terms of frame validity to a universal monadic second-order formula, and thus can be used to express properties of relational structures. An important discipline of modal logic related to this issue—correspondence theory (see van Benthem, 1984)—is mostly about using modal formulas to define classes of relational structures. Thus, as noted in Blackburn et al., 2001, modal languages are simple yet expressive languages for talking about relational structures.

Another reason for interest in some modal logics is that they represent tractable, decidable fragments of first- or second-order logic, which makes them computationally significant. The quest for computational efficiency has recently stimulated active research on the complexity of modal logic.

An important feature of the modal approach is that modal logic presents formal methods of reasoning based on modal operators specific to given practical domains. Often, the linguistic meaning of these operators, coming from their use in the everyday language, is quite imprecise; however, giving the exact semantics for the corresponding logic supplies these modalities with exact meaning. The complete axiomatization with respect to a given formal semantics presents a formal system for reasoning in the corresponding semantic domain, and the completeness theorem with respect to the given interpretation can be considered as a tool for establishing the adequacy of the proposed semantics.

In summary, applied modal logic is a general concept covering modal systems naturally arising from various practical domains. The machinery of applied modal logic contains all tools developed so far in modal logic. Very often the analysis of some new area of application of modal logic needs to invent some new methods, stimulating in this way the general development of the field.

Spatial modal logics. One of the origins of the classical modal logic of *necessity* and *possibility* arises from the analysis of the meaning of these modalities in natural languages. There are many other modes of truth which can be treated as different kinds of modalities: time modalities, space modalities, knowledge modalities, deontic modalities and so on. Examples of time modalities are: *always, sometimes, always in the future, always in the past, at the next moment, tomorrow, since, until*, etc. Examples of space modalities, related to geometrical relations in the space, are: *everywhere, everywhere else, somewhere, somewhere else, near, far, on the left, on the right, on the top, in the middle, between, parallel*, etc. Although geometry, as a mathematical theory of space, is one of the oldest branches of mathematics, and although the theory of time is not even a mathematical discipline, the logic of time is a much

better-established branch of modal logic than the modal logic of space. One explanation of this fact, as noted in Balbiani, 1998, Venema, 1999, is rooted in the use of temporal logic in computer science, especially in program verification and specification, concurrent programming and databases. Another reason is probably in the simpler mathematical structure of time, very suitable for a modal treatment by Kripke semantics: a set of moments of time together with a precedence relation between moments. In contrast, the structure of space is much more complex. For instance, the structure of classical Euclidean geometry consists of several sorts of objects—e.g. points, lines and planes—with various binary relations involving them—e.g. collinearity and betweeness in the set of points, parallelism, concurrence and orthogonality in the set of lines, and the intersort-relations of incidence between different sorts of objects. At first sight, many-sorted mathematical structures are not suitable for modal treatment in the above-described sense, because the standard Kripke semantics is based on one-sorted structures. But as we shall see later, the modal approach has been extended to many-sorted geometric structures.

Recently, in connection with new directions in artificial intelligence and information science, such as geometrical information systems and qualitative spatial reasoning (Cohn and Hazarika, 2001), the application of logic and in particular of modal logic to the theory of space has become more popular, and this has stimulated the development of a new branch of applied modal logic, commonly called *spatial modal logic*.

Lemon and Pratt, 1998 produced a criterion by which one can judge the spatial character of a modal logic and observed in the light of their criterion that several of the existent modal logics of space were not spatial at all. According to Lemon and Pratt, a spatial modal logic is one whose models are based on mathematical models of space. Obviously, affine geometry and projective geometry and some of their fragments constitute mathematical models of space par excellence. Some of these geometrical structures considered as first-order systems are studied in the first part of this chapter. In this second part we include some modal logics with semantics based on them. In this chapter we will neither consider spatial modal logics based on interpretations of modal languages in topology and metric spaces, nor logics based on the primitive notion of a spatial region and some spatial relations between regions like *contact, part-of, overlap* etc., nor modal logics related to the relativistic interpretation of 4-dimensional space-time. The reader can find treatments of such logics in other chapters of this book.

11. Point-based spatial logics

In this section we will consider modal logics related to structures based on a set of points.

The logic of elsewhere and everywhere. One of the first modal logics with explicit spatial interpretation is the logic of "elsewhere" introduced by von Wright, 1979. Under the reading of box given by von Wright, $\Box A$ means "everywhere else it is the case that A". Thus, the modal logic of "elsewhere" may be formally identified with the validities in the class of all Kripke frames $\mathcal{F} = (W, R)$ in which R is the *difference relation*: $\forall x \forall y (xRy \leftrightarrow x \neq y)$. That is why the box and the diamond of von Wright's logic are usually written $[\neq]$ and $\langle \neq \rangle$. This is not a typical spatial relation, because difference can be considered in any set of objects. But what really made von Wright's box popular is the observation that enriching modal languages with $[\neq]$ greatly increases their expressive power, as shown in Goranko, 1990, de Rijke, 1992, Venema, 1993.

The logic of elsewhere can be axiomatized by adding to K the following axioms (von Wright, 1979):

- $A \to [\neq]\langle \neq \rangle A, \quad A \wedge [\neq]A \to [\neq][\neq]A.$

Another example of a simple modal logic with a spatial interpretation is the logic $S5$ considered by Carnap (see Carnap, 1947) as the logic of all structures (W, R) with universal relation R:

$(U) \qquad \forall x, y\colon \quad xRy.$

This gives the following spatial reading of $\Box A$ as "everywhere A".

The condition (U) motivates the box of Carnap's logic to be usually written as $[U]$, and the diamond as $\langle U \rangle$. Note that $[U]A$ is definable in the logic of elsewhere: $[U]A = A \wedge [\neq]A$. Carnap's reading of box has attracted the attention of many logicians, including Goranko and Passy, 1992 and Spaan, 1993.

Collinearity and qualitative distance. Collinearity of points is one of the basic ternary relations between points. Stebletsova, 2000 considers the ternary relation of collinearity between points in projective geometry: $Col(X, Y, Z)$ iff X, Y and Z all lie on a single line. She studies the spatial logic based on Col in any projective geometry of finite dimension $d \geq 2$. The ternary relation Col is used to interpret the binary modality $\circ$ as follows:

$$\mathcal{M}, w \models A \circ B \text{ iff for some } w' \in W \text{ and for some } w'' \in W \text{ with } Col(w, w', w'') \text{ we have } \mathcal{M}, w' \models A \text{ and } \mathcal{M}, w'' \models B.$$

In this setting, the following formulas are valid:

- $A \circ (B \circ C) \to (A \circ B) \circ C,$
- $A \circ B \to B \circ A,$
- $A \to A \circ A$, and

- $\langle U\rangle A \wedge \langle U\rangle B \rightarrow \langle U\rangle(A \circ B)$

where $\langle U\rangle$ is the existential modality between points defined by $\langle U\rangle A = \top \circ A$. Given a finite dimension $d \geq 2$, validity in the class of all projective geometries of dimension d can be axiomatized with a Gabbay-type inference rule (see Gabbay, 1981), but it is not known whether such rules can be replaced by a finite set of additional axioms.

The modal logic of collinearity in projective geometry is rather expressive. For instance, for all finite dimensions $d \geq 2$, there exists a formula in the basic modal language defined above that characterizes exactly those projective spaces of dimension d satisfying the property of Pappus (see Sec. 6.3). Using the fact that Pappus' theorem holds in any finite projective geometry of finite dimension $d \geq 3$, Stebletsova, 2000 has shown that this logic lacks the finite modal property: there exist satisfiable formulas that cannot be satisfied in finite models. What is more, for any finite dimension $d \geq 3$, the satisfiability problem in the class of all projective geometries of dimension d is undecidable. See Stebletsova, 2000 for further details.

Another interesting ternary spatial relation is the relation $N(x, y, z)$ of qualitative distance between points, with the intuitive reading "y is nearer to x than z" (van Benthem, 1983). Its most obvious properties in the real plane may be formulated as follows:

Transitivity $\forall x \forall y \forall z \forall t(N(x, y, z) \wedge N(x, z, t) \rightarrow N(x, y, t))$,

Irreflexivity $\forall x \forall y \neg N(x, y, y)$,

Almost-connectedness $\forall x \forall y \forall z \forall t(N(x, y, z) \rightarrow N(x, y, t) \vee N(x, t, z))$,

Selfishness $\forall x \forall y(x \neq y \rightarrow N(x, x, y))$,

Triangle inequality $\forall x \forall y \forall z(N(x, y, z) \wedge N(z, x, y) \rightarrow N(y, x, z))$.

It is known that N can serve as the basis of elementary plane Euclidean geometry (see Tarski, 1956). Nevertheless, no complete modal spatial logic has been developed so far with Kripke semantics based on that relation (see Aiello and van Benthem, 2002 for further discussion).

12. Line-based spatial logics

In this section we examine some spatial logics based on lines and some standard relations between lines: parallelism, orthogonality and intersection of lines.

The logic of parallelism. Recall that $\parallel$ denotes the relation of strict parallelism. Parallelism frames are structures of the form $(\mathbf{Li}, \parallel)$, where $\mathbf{Li}$ is a non-empty set whose elements are called lines and $\parallel$ is the relation of

strict parallelism between lines. That relation satisfies the following first order conditions:

- $\forall x : x \not\parallel x$ – *no line is parallel to itself*,
- $\forall x, y : x \parallel y$ implies $y \parallel x$ – *the relation* $\parallel$ *is symmetric*,
- $\forall x, y, z : x \parallel y$ and $y \parallel z$ and $x \neq z$ implies $x \parallel z$ – *the relation* $\parallel$ *is "pseudo- transitive"*.

Frames satisfying all these conditions are called strict models of parallelism and their class is denoted by $\mathcal{C}_{SMP}$. The frames satisfying the second and the third axiom are called pre-models of parallelism and their class is denoted by $\mathcal{C}_{PreMP}$.

Balbiani and Goranko, 2002 consider the modal logic of strict parallelism, where $[\|]A$ means *"A is true at all parallel lines"*. The semantics, based on parallelism frames $(\mathbf{Li}, \|)$, is as expected:

$\mathcal{M}, w \models [\|]A$ iff for all $w' \in \mathbf{Li}$ such that $w \parallel w'$, $\mathcal{M}, w' \models A$.

Obviously, the following formulas modally define the class $\mathcal{C}_{PreMP}$:

$$A \to [\|]\langle\|\rangle A, \quad A \wedge [\|]A \to [\|][\|]A.$$

We denote by PAR the axiom system obtained by adding these formulas to the minimal modal logic K. Let us note that these axioms are of Sahlqvist type and just modally define the class of frames $\mathcal{C}_{PreMP}$. Then, by the Sahlqvist theorem, PAR is sound and complete with respect to $\mathcal{C}_{PreMP}$.

Let us remark that PAR and the logic of elsewhere are the same. Hence, repeating the completeness proof of the logic of elsewhere given in Segerberg, 1981, one can show that PAR is also complete with respect to the strict models of parallelism $\mathcal{C}_{SMP}$. The difficulty of working with strict models of parallelism is that there is no formula corresponding to the irreflexivity of the relation $\|$. This lack of expressive power of the modal language enables us to show that the satisfiability problem $SAT(\mathcal{C}_{SMP})$ is NP-complete (Demri, 1996).

Adding to PAR the following Sahlqvist formulas for all $n \geq 0$, we obtain the axiom system PAR^E:

(φ_0) $\langle\|\rangle\top$,

(φ_n) $\langle\|\rangle[\|]A_1 \wedge \ldots \wedge \langle\|\rangle[\|]A_n \to \langle\|\rangle(A_1 \wedge \ldots \wedge A_n)$.

The formula φ_n corresponds to the following first-order property on parallelism frames:

(Φ_n) $x \parallel y_1 \wedge \ldots \wedge x \parallel y_n \to (\exists z)(x \parallel z \wedge y_1 \parallel z \wedge \ldots \wedge y_n \parallel z)$.

Note that φ_n is derivable from φ_{n+1}.

Since the model $\mathcal{F}_{n+2} = (\{1, \ldots, n+2\}, \neq)$ of strict parallelism consisting of exactly $n+2$ parallel lines validates φ_n but does not validate φ_{n+1}, we infer that PAR^E is not finitely axiomatizable. Nevertheless, since φ_n is a Sahlqvist formula for all $n \geq 0$, PAR^E is sound and complete with respect to the class $\mathcal{C}^{\infty}_{PreMP}$ of all pre-models of parallelism satisfying (Φ_n), for all $n = 0, 1, \ldots$. Repeating the line of reasoning suggested by Segerberg, 1981 within the context of the logic of elsewhere, one can show that PAR^E is also complete with respect to the class $\mathcal{C}^{\infty}_{SMP}$ consisting of all strict models of parallelism satisfying the conditions (Φ_n) for all $n = 0, 1, \ldots$.

A more interesting completeness result for the logic PAR^E is that it is sound and complete both in the standard parallelism frames in real plane $\mathbb{P}^2$ and in real 3-dimensional space $\mathbb{P}^3$ with the usual relation of strict parallelism. This shows that the language of strict parallelism is not expressive enough to distinguish the standard parallelism frames in $\mathcal{C}^{\infty}_{PreMP}$. Despite this obvious lack of expressive power of our modal language, the advantage of our modal approach is that the decision problem $SAT(\mathcal{C}^{\infty}_{SMP})$ for satisfiability in $\mathcal{C}^{\infty}_{SMP}$ is also NP-complete (see Balbiani and Goranko, 2002 for the details).

The logic of orthogonality. The relation of orthogonality $\perp$ is another typical binary relation between lines. We interpret $[\perp]A$ as "*A is true at all lines which lie orthogonal to the current line*". Let us note that in every orthogonality frame $\mathcal{F} = (\mathbf{Li}, \perp)$, the binary relation $\|$, defined as follows, is an equivalence relation:

$$w \parallel w' \text{ iff for all lines } w'', w \perp w'' \text{ iff } w' \perp w''.$$

We consider the class $\mathcal{C}_{PQMO}$ of all planar quasi-models of orthogonality $\mathcal{F} = (\mathbf{Li}, \perp)$, where $\perp$ is symmetric and 3-transitive, i.e.:

- $\forall w \forall w' (w \perp w' \rightarrow w' \perp w)$ – symmetry of $\perp$,
- $\forall w \forall w' \forall w'' \forall w''' (w \perp w' \wedge w' \perp w'' \wedge w'' \perp w''' \rightarrow w \perp w''')$ – 3-transitivity of $\perp$.

The class $\mathcal{C}_{PMLO}$ of *planar models of line orthogonality* is the class of all frames $\mathcal{F} = (\mathbf{Li}, \perp)$ where $\perp$ is irreflexive, symmetric and 3-transitive. Let $\mathcal{C}^{\infty}_{PQMO}$ be the class of all planar quasi-models of line orthogonality in which every equivalence class modulo $\|$ is infinite. Similarly, let $\mathcal{C}^{\infty}_{PMLO}$ be the class of all planar models of line orthogonality in which every equivalence class modulo $\|$ is infinite.

The modal logic ORT based on $\mathcal{C}_{PQMO}$ is obtained by adding the following axioms to K:

- $A \rightarrow [\perp]\langle\perp\rangle A, \quad [\perp]A \rightarrow [\perp][\perp][\perp]A.$

These axioms are formulas of Sahlqvist type just defining the properties of symmetry and 3-transitivity of the relation $\perp$, so by the Sahlqvist theorem ORT is complete in $\mathcal{C}_{PQMO}$. Using bounded morphisms one may prove that the classes $\mathcal{C}_{PQMO}$ and $\mathcal{C}_{PMLO}$ define the same logic, which shows that ORT is also complete in the class $\mathcal{C}_{PMLO}$. See Balbiani and Goranko, 2002 for details.

The logic which corresponds to the class of frames $\mathcal{C}^{\infty}_{PQMO}$ is finitely axiomatizable through the axiom system ORT^E obtained by adding to ORT the axiom $\langle\perp\rangle\top$. This axiom system is sound and complete also with respect to validity in the Euclidean orthogonality plane consisting of all lines in the real plane together with the usual orthogonality relation (see Balbiani and Goranko, 2002). So the language of orthogonality is not expressive enough to distinguish the standard orthogonality frame in the class $\mathcal{C}^{\infty}_{PQMO}$.

Let us note that the formula $[\perp]A \to [\perp][\perp][\perp]A$ is not valid in the Euclidean orthogonality space consisting of all lines in real 3-dimensional space together with the usual orthogonality relation. Hence, comparing with the modal logic of parallelism, the modal logic of orthogonality is able to distinguish the Euclidean orthogonality plane and the Euclidean orthogonality 3-dimensional space.

A modal logic of parallelism and intersection of lines. Having outlined the modal logics of parallelism and the modal logics of orthogonality, we are now in a position to consider richer geometrical structures. Specifically, we discuss the line-based modal logic based on the binary relations of parallelism and intersection of lines, with the corresponding modal operators $[\|]$ and $[\times]$. Our aim is to axiomatize the logic of the standard two-dimensional frame consisting of all lines in the real affine plane (called SAP) with the strict parallelism relation $\|$ and the standard relation of intersection: $a \times b$ iff a and b have only one common point. Let us note that the following modal formulas are true in SAP:

- $\varphi \to [\|]\langle\|\rangle\varphi$,
- $\varphi \wedge [\|]\varphi \to [\|][\|]\varphi$,
- $\varphi \to [\times]\langle\times\rangle\varphi$
- $[\times]\varphi \to [\|][\times]\varphi$,
- $\varphi \wedge [\|]\varphi \wedge [\times]\varphi \to [\times][\times]\varphi$,
- $\langle\|\rangle\top$,
- $\langle\|\rangle\varphi_1 \wedge \ldots \wedge \langle\|\rangle\varphi_n \to \langle\|\rangle(\langle\|\rangle\varphi_1 \wedge \ldots \wedge \langle\|\rangle\varphi_n)$, $n = 1, 2, \ldots,$
- $\langle\times\rangle\top$,

- $\langle\times\rangle\varphi_1 \wedge \ldots \wedge \langle\times\rangle\varphi_n \to \langle\times\rangle(\langle\times\rangle\varphi_1 \wedge \ldots \wedge \langle\times\rangle\varphi_n), n = 1, 2, \ldots$.

We denote by $\mathcal{ML}(SAP)$ the extension of the logic K with these axioms. Since all of them are Sahlqvist formulas, they define the following class $\mathcal{C}(PreSAP)$ of frames (called pre-standard affine planes) in which $\mathcal{ML}(SAP)$ is complete:

- $u \parallel v \to v \parallel u$,
- $u \parallel v \wedge v \parallel w \wedge u \neq w \to u \parallel w$,
- $u \times v \to v \times u$,
- $u \parallel v \wedge v \times w \to u \times w$,
- $u \times v \wedge v \times w \to u = w \vee u \parallel w \vee u \times w$,
- $(\forall u \exists v)(u \parallel v)$,
- $u \parallel v_1 \wedge \ldots \wedge u \parallel v_n \to (\exists w)(u \parallel w \wedge w \parallel v_1 \wedge \ldots \wedge w \parallel v_n)$, $n = 1, 2, \ldots$,
- $(\forall u \exists v)(u \times v)$,
- $u \times v_1 \wedge \ldots \wedge u \times v_n \to (\exists w)(u \times w \wedge w \times v_1 \wedge \ldots \wedge w \times v_n), n = 1, 2, \ldots$.

Let us call a structure $(\mathbf{Li}, \parallel, \times)$ a *general affine plane* if it satisfies all of the above first-order conditions plus the conditions of irreflexivity of the relations $\times$ and $\parallel$, and let us denote the class of all such structures by $\mathcal{C}(GAP)$. Note that the standard affine plane SAP is in this class. Applying the method of bounded morphisms it can be proved that $\mathcal{ML}(SAP)$ is also complete in this class. However this still does not prove that $\mathcal{ML}(SAP)$ is complete with respect to SAP. Applying more complicated techniques from model theory, it can be proved that any two frames from $\mathcal{C}(GAP)$ are modally equivalent, i.e. determine equal logics. Since the standard affine plane is in $\mathcal{C}(GAP)$, this implies that the logic $\mathcal{ML}(SAP)$ is complete for its standard semantics.

The classes $\mathcal{C}(PreSAP)$ and $\mathcal{C}(GAP)$ are quite different. Using the selective filtration techniques one can prove that the logic $\mathcal{ML}(SAP)$ has finite model property (fmp) with respect to $\mathcal{C}(PreSAP)$ and hence is decidable, while with respect to $\mathcal{C}(GAP)$ it does not have fmp – all frames from $\mathcal{C}(GAP)$ are infinite. These facts, however, help to prove that satisfiability problem for $\mathcal{C}(GAP)$ is NP-complete.

The completeness theorem for $\mathcal{ML}(SAP)$ implies that the modal language of parallelism and intersection of lines is too weak to distinguish the standard 2-dimensional frame from the other frames in the class $\mathcal{C}(GAP)$. But the language can distinguish the 2-dimensional standard frame from the 3-dimensional

standard frame, the later consisting of all lines in real 3-dimensional space with the standard relation of strict parallelism and the standard relation of intersection of lines. For instance the following axiom of $\mathcal{ML}(SAP)$ is not true in 3-dimensional space: $[\times]A \rightarrow [\|][\times]A$. The reason is that in 3-dimensional space there are lines intersecting one of two parallel lines but not the other.

13. Tip spatial logics

Projective geometry and affine geometry are among the most prominent mathematical models of space. They arise from the study of points and lines by means of properties stated in terms of incidence. In this section and in the following one, we will introduce modal logics for incidence between points and lines. There are two different approaches for defining a modal logic of incidence between points and lines. The standard semantics for modal logic assumes Kripke models with only one sort of possible worlds. Therefore, the first approach consists in the replacement of the two-sorted structures based on points and lines by one-sorted structures containing the same geometrical information. The second approach consists in the extension of the modal logic formalism allowing two sorts of formulas, point formulas and line formulas, and two sorts of possible worlds in Kripke models. The remainder of this section briefly describes the first approach (see Balbiani et al., 1997), while the second approach will be considered in Sec. 14.

Tips. Let $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$ be a point-line incidence plane, that is:

- $\mathbf{Po}$ is a non-empty set of points with typical elements denoted by X, Y, Z, T, etc, possibly with subscripts,
- $\mathbf{Li}$ is a non-empty set of lines with typical elements denoted by x, y, z, t, etc, possibly with subscripts,
- I is a binary relation of incidence between points and lines.

The relationship XIx will be read "X is incident with x", "X lies in x", "x is incident with X", or "x passes through X". We will always assume that $\mathbf{Po} \cap \mathbf{Li} = \emptyset$, i.e. no point is a line and no line is a point. Hereafter, we will assume in this section that the binary relation I satisfies the following first-order conditions:

- $\forall X \forall Y \exists z (XIz \wedge YIz)$,
- $\forall X \forall Y \forall z \forall t (XIz \wedge YIz \wedge XIt \wedge YIt \rightarrow X = Y \vee z = t)$,
- $\forall x \exists Y \exists Z (YIx \wedge ZIx \wedge Y \neq Z)$,
- $\forall X \exists y \exists z (XIy \wedge XIz \wedge y \neq z)$.

The notion of point-line incidence plane can be extended with new first-order conditions in different directions. Two natural extensions are the notion of affine plane and the notion of projective plane. Consider a point-line incidence plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$. Let us define on $\mathbf{Li}$ the binary relation $\parallel$ in the following way:

- $x \parallel y$ iff for all points Z, if ZIx and ZIy then $x = y$,

A point-line incidence plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$ is called an affine plane if it satisfies the following additional first-order conditions:

- $\forall X \forall y \exists z (XIz \wedge y \parallel z)$, $\forall x \forall y \forall z (x \parallel y \wedge y \parallel z \rightarrow x \parallel z)$.

Obviously, point-line affine planes are Euclidean in the sense that they satisfy the following condition:

- $\forall X \forall y \forall z \forall t (XIz \wedge y \parallel z \wedge XIt \wedge y \parallel t \rightarrow z = t)$.

A point-line incidence plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$ is called a projective plane if it satisfies the following additional first-order conditions:

- $\forall x \forall y \exists Z (ZIx \wedge ZIy)$,
- $\forall x \forall Y \forall Z \exists T (YIx \wedge ZIx \rightarrow TIx \wedge T \neq Y \wedge T \neq Z)$.

It is clear from our definition that if $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$ is a projective plane then two different points are always incident with exactly one line whereas two different lines have always one point in common.

Traditionally, the Kripke semantics of modal logics is based on one-sorted relational structures. That is why we introduce a new kind of relational structures, called incidence frames, which are one-sorted and which will be used for defining the Kripke semantics of our next spatial logics. Now consider a point-line incidence plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$. We shall say that the pair (X, x) in $\mathbf{Po} \times \mathbf{Li}$ is a *tip* over $\mathcal{F}$ iff XIx. Intuitively, the tip (X, x) can be considered both as the point X and as the line x. Using tips, we can define the following binary relations:

- $(X, x) \equiv_1^{\mathcal{F}} (Y, y)$ iff $X = Y$,
- $(X, x) \equiv_2^{\mathcal{F}} (Y, y)$ iff $x = y$.

In the expression $(X, x) \equiv_1^{\mathcal{F}} (Y, y)$, (X, x) and (Y, y) are considered as the points X and Y and the relation $\equiv_1^{\mathcal{F}}$ can be seen as the equality of points. Similarly, in the expression $(X, x) \equiv_2^{\mathcal{F}} (Y, y)$, (X, x) and (Y, y) are considered as the lines x and y and the relation $\equiv_2^{\mathcal{F}}$ can be seen as the equality of lines.

Using the binary relations $\equiv_1^{\mathcal{F}}$ and $\equiv_2^{\mathcal{F}}$, we can simulate the binary relation of incidence between points and lines and the binary relation of parallelism

between lines in $\mathcal{F}$. Let $O^{\mathcal{F}}$ and $\|^{\mathcal{F}}$ be the binary relations between tips defined in the following way:

- $(X, x)O^{\mathcal{F}}(Y, y)$ iff XIy,
- $(X, x) \|^{\mathcal{F}} (Y, y)$ iff $x \| y$.

Obviously, the relation of incidence $O^{\mathcal{F}}$ between tips is definable by means of $\equiv_1^{\mathcal{F}}$ and $\equiv_2^{\mathcal{F}}$ as follows: $w_1 O^{\mathcal{F}} w_2$ iff there exists a tip w such that $w_1 \equiv_1^{\mathcal{F}} w$ and $w \equiv_2^{\mathcal{F}} w_2$, hence $O^{\mathcal{F}} = \equiv_1^{\mathcal{F}} \circ \equiv_2^{\mathcal{F}}$ where $\circ$ is the composition of binary relations. Likewise, $w_1 \|^{\mathcal{F}} w_2$ iff for all tips w, if $wO^{\mathcal{F}}w_1$ and $wO^{\mathcal{F}}w_2$ then $w_1 \equiv_2^{\mathcal{F}} w_2$.

Incidence frames. Tips motivate the following definition. Consider a point-line incidence plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$. The incidence frame over $\mathcal{F}$ is the structure $W(\mathcal{F}) = (W^{\mathcal{F}}, \equiv_1^{\mathcal{F}}, \equiv_2^{\mathcal{F}})$ where $W^{\mathcal{F}}$ is the set of all tips over $\mathcal{F}$. It is not too difficult to see that $\equiv_1^{\mathcal{F}}$ and $\equiv_2^{\mathcal{F}}$ are equivalence relations on W satisfying the following additional conditions:

$(I1)$ $\forall w \forall w' (w \equiv_1^{\mathcal{F}} w' \wedge w \equiv_2^{\mathcal{F}} w' \rightarrow w = w')$,

$(I2)$ $\forall w \forall w' \exists w'' (wO^{\mathcal{F}}w'' \wedge w'O^{\mathcal{F}}w'')$,

$(I3)$ $\forall w \forall w' \forall w'' \forall w''' (wO^{\mathcal{F}}w'' \wedge w'O^{\mathcal{F}}w'' \wedge wO^{\mathcal{F}}w''' \wedge w'O^{\mathcal{F}}w''' \rightarrow$
$w \equiv_1^{\mathcal{F}} w' \vee w'' \equiv_2^{\mathcal{F}} w''')$,

$(I4)$ $\forall w \exists w' \exists w'' (w'O^{\mathcal{F}}w \wedge w''O^{\mathcal{F}}w \wedge w' \not\equiv_1^{\mathcal{F}} w'')$,

$(I5)$ $\forall w \exists w' \exists w'' (wO^{\mathcal{F}}w' \wedge wO^{\mathcal{F}}w'' \wedge w' \not\equiv_2^{\mathcal{F}} w'')$.

Let us remark that $\equiv_1$ and $\equiv_2$ define $=$ in the following way: $w = w'$ iff $w \equiv_1 w'$ and $w \equiv_2 w'$. Moreover, if $\mathcal{F}$ is affine then:

$(A1)$ $\forall w \forall w' \exists w'' (wO^{\mathcal{F}}w'' \wedge w' \|^{\mathcal{F}} w'')$,

$(A2)$ $\forall w \forall w' \forall w'' (w \|^{\mathcal{F}} w' \wedge w' \|^{\mathcal{F}} w'' \rightarrow w \|^{\mathcal{F}} w'')$.

If $\mathcal{F}$ is projective then:

$(P1)$ $\forall w \forall w' \exists w'' (w''O^{\mathcal{F}}w \wedge w''O^{\mathcal{F}}w')$,

$(P2)$ $\forall w \forall w' \forall w'' \exists w''' (w'O^{\mathcal{F}}w \wedge w''O^{\mathcal{F}}w \rightarrow$
$w'''O^{\mathcal{F}}w \wedge w''' \not\equiv_1^{\mathcal{F}} w' \wedge w''' \not\equiv_1^{\mathcal{F}} w'')$.

These conditions are characteristic in the following sense: if in a set W we have two equivalence relations $\equiv_1$ and $\equiv_2$ satisfying the conditions $(I1)$–$(I5)$ then there exists a point-line incidence plane $\mathcal{F}$ such that the relational structures $(W, \equiv_1, \equiv_2)$ and $W(\mathcal{F}) = (W^{\mathcal{F}}, \equiv_1^{\mathcal{F}}, \equiv_2^{\mathcal{F}})$ are isomorphic. Moreover,

if $(W, \equiv_1, \equiv_2)$ satisfies the conditions $(A1)$ and $(A2)$ then the corresponding point-line incidence plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$ is affine, and, if it satisfies the conditions $(P1)$ and $(P2)$ then the corresponding point-line incidence plane is projective. So, in order to define the Kripke semantics of a modal logic of incidence, we can use one-sorted structures of the form $(W, \equiv_1, \equiv_2)$ instead of point-line incidence planes.

Consider a relational structure of the form $(W, \equiv_1, \equiv_2)$ where $\equiv_1$ and $\equiv_2$ are equivalence relations on W. We shall say that $(W, \equiv_1, \equiv_2)$ is an incidence frame if it satisfies the conditions $(I1)$–$(I5)$. Moreover, $(W, \equiv_1, \equiv_2)$ is said to be affine if it satisfies $(A1)$ and $(A2)$, and it is said to be projective if it satisfies $(P1)$ and $(P2)$.

Let us note that the properties of Desargues and Pappus are also expressible in the present language, so it is quite rich.

A modal logic for incidence. Our modal language for incidence frame uses the modal operators $[\equiv_1]$, $[\equiv_2]$, $[\not\equiv_1]$, and $[\not\equiv_2]$. Well-formed formulas are given by the rule:

- $A ::= p \mid \bot \mid \neg A \mid (A \vee B) \mid [\equiv_1]A \mid [\equiv_2]A \mid [\not\equiv_1]A \mid [\not\equiv_2]A.$

Abbreviations: difference – $[\neq]A = [\not\equiv_1]A \wedge [\not\equiv_2]A$, universal modality – $[U]A = A \wedge [\neq]A$, incidence $[O]A = [\equiv_1][\equiv_2]A$, $[O^{-1}]A = [\equiv_2][\equiv_1]A$.

The semantics is based on incidence frames in the expected way. In particular, we have:

- $\mathcal{M}, w \models [\equiv_i]A$ iff for all $w' \in W$ such that $w \equiv_i w'$, $\mathcal{M}, w' \models A$,
- $\mathcal{M}, w \models [\not\equiv_i]A$ iff for all $w' \in W$ such that $w \not\equiv_i w'$, $\mathcal{M}, w' \models A$,

for $i \in \{1, 2\}$.

The following formulas are valid in $\mathcal{C}_{inc}$:

(Ax_1) $A \to [\not\equiv_i]\langle\not\equiv_i\rangle A$, $i \in \{1, 2\}$,

(Ax_2) $A \to [\neq]\langle\neq\rangle A$,

(Ax_3) $A \wedge [\neq]A \to [\neq][\neq]A$,

(Ax_4) $[U]A \to [\equiv_i]A$, $i \in \{1, 2\}$,

(Ax_5) $[\equiv_i]A \wedge [\not\equiv_i]A \to [U]A$, $i \in \{1, 2\}$,

(Ax_6) $\langle\not\equiv_i\rangle A \to [\equiv_i]\langle\neq\rangle A$, $i \in \{1, 2\}$,

(Ax_7) $[\equiv_i]A \to A$, $A \to [\equiv_i]\langle\equiv_i\rangle A$, $[\equiv_i]A \to [\equiv_i][\equiv_i]A$, $i \in \{1, 2\}$,

(Ax_8) $[O][O^{-1}]A \to [U]A$,

(Ax_9) $\langle O\rangle(A \wedge \langle O^{-1}\rangle([\neq]B \wedge C)) \rightarrow ([O](\langle\equiv_2\rangle A \vee [O^{-1}]B) \vee \langle\equiv_1\rangle C)$,

(Ax_{10}) $A \rightarrow \langle O^{-1}\rangle\langle\not\equiv_1\rangle\langle O\rangle A$,

(Ax_{11}) $A \rightarrow \langle O\rangle\langle\not\equiv_2\rangle\langle O^{-1}\rangle A$.

Let MIG (Modal Incidence Geometry) be the axiom system obtained by adding the formulas (Ax_1)–(Ax_{11}) to the minimal normal modal logic in our language. Note that all proper axioms of MIG are Sahlqvist formulas and that the associated first-order properties correspond to conditions defining incidence frames. For example, the formulas (Ax_9), (Ax_{10}), and (Ax_{11}) correspond to the first-order properties $(I3)$, $(I4)$, and $(I5)$ respectively. Nevertheless, MIG is not known to be complete with respect to validity in the class $\mathcal{C}_{inc}$ of all incidence frames. The point is that the interpretation in incidence frames of formulas in the form $[\not\equiv_1]A$ and $[\not\equiv_2]A$ is based on the complements $\not\equiv_1$ and $\not\equiv_2$ of the binary relations $\equiv_1$ and $\equiv_2$. The difficulty with the complementarity relations is that there is no axiom corresponding exactly to the first-order properties saying that:

- $\equiv_i \cap \not\equiv_i = \emptyset$ for $i \in \{1, 2\}$.

We have seen that $[\not\equiv_1]$ and $[\not\equiv_2]$ define $[\neq]$. Moreover, notice that on the class $\mathcal{C}_{inc}$ of all incidence frames the formula $A \wedge [\neq]\neg A$ is satisfied at some tip w in some incidence model $\mathcal{M} = (W, \equiv_1, \equiv_2, V)$ iff w is the only tip in W where A holds. Hence, $A \wedge [\neq]\neg A$ can be considered as a sort of proper name for w. The reader may observe that the first-order properties saying that the binary relations $\equiv_i$ and $\not\equiv_i$ are disjoint are equivalent to the first-order condition of irreflexivity of the binary relation $\not\equiv_1 \cup \not\equiv_2$. Although irreflexivity does not correspond to a modal formula, it can be characterized in some sense by an inference rule. In this connection, see Gabbay, 1981, de Rijke, 1992, and Venema, 1993. This suggests that we enrich the axiom system MIG with a special inference rule, the inference rule of irreflexivity:

- "Given $p \wedge [\neq]\neg p \rightarrow A$, prove A",

where p is a proposition letter not occurring in A, thus obtaining the axiom system MIG^+. This inference rule has also an infinitary version:

- "Given $p \wedge [\neq]\neg p \rightarrow A$ for all proposition letters, prove A",

which gives rise to the same set of provable formulas. Soundness of MIG^+ with respect to validity in $\mathcal{C}_{inc}$ is straightforward: we already know that MIG is sound, hence, it is enough to verify that the inference rule of irreflexivity preserves validity in the class $\mathcal{C}_{inc}$. As for the completeness of MIG^+, we build a special model from maximal consistent sets of formulas which are closed under the infinitary version of the rule. Since all proper axioms of MIG^+ are Sahlqvist formulas and our modal language is versatile, the underlying relational

structure is an incidence frame. See Blackburn et al., 2001; Venema, 1993 for more details about the importance of inference rules like the inference rule of irreflexivity. We do not know if it is possible to eliminate the inference rule of irreflexivity in our axiom system: the completeness of MIG with respect to validity in the class $\mathcal{C}_{inc}$ is still open. In addition, the decidability/complexity issue of validity in $\mathcal{C}_{inc}$ is still unresolved.

Note also that we can obtain a complete axiomatization of the projective incidence frames adding the following axioms to the system MIG:

$(\mathbf{MPG1})$ $\langle U\rangle A \to \langle O^{-1}\rangle\langle O\rangle A$,

$(\mathbf{MPG2})$ $\langle O\rangle(A \wedge \langle O^{-1}\rangle B) \to \langle \not\equiv\rangle(\langle O\rangle A \wedge \langle \not\equiv\rangle B)$.

Extending the language with the modality $[\|]$ we can axiomatize also the affine incidence frames.

We note that the presented systems have rich expressiveness, containing modalities with the following intuitive readings: $[U]A$ – everywhere; $[\neq]A$ – everywhere else; $[\equiv_1]A$ – in all points; $[\not\equiv_1]A$ – in all other points; $[\equiv_2]A$ – in all lines; $[\not\equiv_2]A$ – in all other lines; $[O]A$ – in all lines through the current point; $[O^{-1}]A$ – in all points on the current line.

14. Point-line spatial logics

Standard modal languages have semantics over one-sorted frames. Within the context of dynamic logic, van Benthem, 1994, Marx, 1996, and de Rijke, 1995 were among the first to use relational structures made up of several sets of possible worlds together with binary relations between them. One possible application of such languages are many-sorted geometrical structures like incidence geometries based on points and lines and inter-sort relations of incidence between them. In this section we follow Venema, 1999.

Two-sorted modal logic. Consider, for instance, a relational structure of the form $\mathcal{F} = (W_1, W_2, R)$ where W_1 and W_2 are nonempty sets and $R \subseteq W_1 \times W_2$. For the sake of simplicity, we assume that the sets W_1 and W_2 are disjoint. From now on, such structures will be called two-sorted Kripke frames. In $\mathcal{F}$, the binary relation R links elements of W_1 with elements of W_2. If modal languages must be used for talking about relational structures like $\mathcal{F}$, one possibility is to consider a language with two sorts of formulas:

- $A ::= p \mid \bot \mid \neg A \mid (A \vee B) \mid \Box\alpha$ – *formulas of the first sort*,
- $\alpha ::= \pi \mid \bot \mid \neg\alpha \mid (\alpha \vee \beta) \mid \Box A$ – *formulas of the second sort*.

where p and π denote propositional letters of the corresponding sorts. Note that the modality $\Box$ transform the one sort into the other.

A two-sorted Kripke model based on $\mathcal{F}$ is nothing but a structure of the form $\mathcal{M} = (W_1, W_2, R, V)$ where V—the valuation of the model—associates a subset $V(p)$ of W_1 with every propositional letter p of the first type and a subset $V(\pi)$ of W_2 with every propositional letter π of the second type. Elements of W_1 being denoted by upper case letters like X, Y, Z, etc, and elements of W_2 being denoted by lower case letters like x, y, z, etc, formulas like p, $\neg A$, $A \vee B$, and $\Box\alpha$ will be interpreted at elements of W_1, whereas formulas like π, $\neg\alpha$, $\alpha \vee \beta$, and $\Box A$ will be interpreted at elements of W_2 according to the satisfiability relation defined as usual. In particular:

- $\mathcal{M}, X \models \Box\alpha$ iff $(\forall y \in W_2)(XRy$ implies $\mathcal{M}, y \models \alpha)$,
- $\mathcal{M}, x \models \Box A$ iff $(\forall Y \in W_1)(YRx$ implies $\mathcal{M}, Y \models A)$.

We can define what it means for a formula of a given sort to be true (satisfiable) in given model in the obvious way.

Although the extension of the standard techniques (canonical model, bisimulation, filtration, etc) and results (completeness, finite frame property, definability, etc) of modal logic to multi-sorted languages like the one we have just described has never been considered in detail we believe that their extension to multi-sorted modal logics is straightforward. As for the two-sorted modal language considered above, it is a simple matter to check that K_2—the following axiom system—is sound and complete with respect to validity in the class of all two-sorted Kripke frames:

- Axioms of the first type: all first-type substitution instances of classical tautologies together with all formulas of the form
  $\Box(\alpha \to \beta) \to (\Box\alpha \to \Box\beta)$ and $A \to \Box\Diamond A$,
- Axioms of the second type: all second-type substitution instances of classical tautologies together with all formulas of the form
  $\Box(A \to B) \to (\Box A \to \Box B)$ and $\alpha \to \Box\Diamond\alpha$,
- Inference rules of the first type: Modus ponens "given A and $A \to B$, prove B" and generalization "given α, prove $\Box\alpha$",
- Inference rules of the second type: Modus ponens "given α and $\alpha \to \beta$, prove β" and generalization "given A, prove $\Box A$".

In the next paragraph we apply these ideas to the cases of plane projective geometry and plane affine geometry.

A two-sorted modal logic for plane projective geometry. The point-line incidence planes defined in Sec. 13 are good candidates to stand for the Kripke semantics of a two-sorted modal language. Let us consider the class $\mathcal{C}_{pg}$ of all

projective planes, i.e. two-sorted structures $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$ where $\mathbf{Po}$ is a nonempty set of points, $\mathbf{Li}$ is a nonempty set of lines, and I is a binary relation between points and lines such that:

- $\forall X \forall Y \exists z (XIz \wedge YIz)$,
- $\forall x \forall y \exists Z (ZIx \wedge ZIy)$,
- $\forall X \forall Y \forall z \forall t (XIz \wedge YIz \wedge XIt \wedge YIt \rightarrow X = Y \vee z = t)$.

Of course, we will assume that the sets $\mathbf{Po}$ and $\mathbf{Li}$ are disjoint. According to the discussion above, we now turn to the definition of our two-sorted modal languages for talking about projective planes. Let us consider a countable set $\Phi_{\mathbf{Po}}$ of point-type proposition letters, with typical members denoted p, q, r, etc, and a countable set $\Phi_{\mathbf{Li}}$ of line-type proposition letters, with typical members denoted π, ρ, σ, etc. The well-formed formulas are defined by the following rules:

- $A ::= p \mid \bot \mid \neg A \mid (A \vee B) \mid \Box\alpha$ – *point formulas*,
- $\alpha ::= \pi \mid \bot \mid \neg\alpha \mid (\alpha \vee \beta) \mid \Box A$ – *line formulas*.

In some two-sorted model $\mathcal{M} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I}, V)$, the point-formula $\Box\alpha$ is satisfied at a point X iff the line-formula α is satisfied at every line passing through X. Similarly, the line-formula $\Box A$ is satisfied at line x iff the point-formula A is satisfied at every point lying on x.

Seeing that two points are always incident with at least one line and two lines are passing together through at least one point, the reader may easily verify that the point-formula $\Box\Box A$ is satisfied at point X in $\mathcal{M}$ iff the point-formula A is true everywhere in $\mathcal{M}$. Similarly, the line-formula $\Box\Box\alpha$ is satisfied at line x in $\mathcal{M}$ iff the line-formula α is true everywhere in $\mathcal{M}$. Hence, the universal modality $[U]$ for points and the universal modality $[u]$ for lines are definable in the following way: $[U]A = \Box\Box A$ and $[u]\alpha = \Box\Box\alpha$.

The two-sorted modal logic defined by the class $\mathcal{C}_{pg}$ of all projective planes has been studied first by Balbiani, 1998 and Venema, 1999. They proved that the axiom system K_{pg} obtained by adding all instances of the following axioms to K_2 is complete with respect to validity in $\mathcal{C}_{pg}$:

Axioms of the first type:	Axioms of the second type:
$\Box\alpha \rightarrow [U]\Diamond\alpha$	$\Box A \rightarrow [u]\Diamond A$
$[U]A \rightarrow A$	$[u]\alpha \rightarrow \alpha$
$[U]A \rightarrow [U][U]A$	$[u]\alpha \rightarrow [u][u]\alpha$
$A \rightarrow [U]\langle U\rangle A$	$\alpha \rightarrow [u]\langle u\rangle\alpha$

The proof of the decidability of the set of all formulas of the first type satisfiable in $\mathcal{C}_{pg}$ and the proof of the decidability of the set of all formulas of

the second type satisfiable in $\mathcal{C}_{pg}$ can be done using the standard technique of the filtration. As usual, this filtration argument implies that satisfiability of point-formulas and line-formulas within $\mathcal{C}_{pg}$ is in NEXPTIME. What makes interesting our two-sorted modal logic is the following result proved by Venema, 1999: satisfiability of point-formulas and line-formulas within the class $\mathcal{C}_{pg}$ is NEXPTIME-complete.

The expressive power of our two-sorted modal language is weak. For example, neither the difference modality between points nor the difference modality between lines are definable in it. Let us extend our two-sorted language by allowing point-formulas like $[D]A$ and line-formulas like $[d]\alpha$. In some two-sorted model $\mathcal{M} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I}, V)$, the point-formula $[D]A$ will be satisfied at point X iff point-formula A is satisfied at every point different from X whereas the line-formula like $[d]\alpha$ will be satisfied at line x iff line-formula α is satisfied at every line different from x. The axiomatisation/completeness and decidability/complexity issues of validity and satisfiability of formulas in the extended two-sorted language with respect to the class $\mathcal{C}_{pg}$ are still open.

A two-sorted modal logic for plane affine geometry. A particular aspect of plane projective geometry is the duality between points and lines. In plane affine geometry, points and lines are no longer interchangeable seeing that, in point-line affine planes, although two different points are always incident with exactly one line, parallel lines have no point in common. This imbalance between points and lines in affine planes is translated into additional difficulties for those who wish to define a two-sorted modal logic for plane affine geometry. The language of this two-sorted modal logic must be able to talk about incidence between points and lines and parallelism between lines. The solution in Balbiani and Goranko, 2002 is to consider the following rules that mutually define the formulas of sort point and the formulas of sort line:

- $A ::= p \mid \bot \mid \neg A \mid (A \vee B) \mid \Box\alpha$,
- $\alpha ::= \pi \mid \bot \mid \neg\alpha \mid (\alpha \vee \beta) \mid \Box A \mid [\|_s]\alpha$.

As for the two-sorted modal logic for projective geometry, point formulas like $\Box\alpha$ are read "α is satisfied at every line incident with the current point" and line formulas like $\Box A$ are read "A is satisfied at every point incident with the current line". The unary modality $[\|_s]$ will be interpreted by the strong, i.e. irreflexive, binary relation of parallelism between lines defined, in any affine plane $\mathcal{F} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I})$, by:

- $x \|_s y$ iff for all points Z, not ZIx or not ZIy.

In this setting, if V is a valuation on $\mathcal{F}$ then the definition of the satisfiability relation in the two-sorted model $\mathcal{M} = (\mathbf{Po}, \mathbf{Li}, \mathbf{I}, V)$ defined by V over $\mathcal{F}$ now contains the following item:

- $\mathcal{M}, x \models [\|_s]\alpha$ iff for all $y \in \mathbf{Li}$ such that $x \|_s y$, $\mathcal{M}, y \models \alpha$.

It is a simple matter to check that, in $\mathcal{M}$, the universal modality $[U]$ for points and the universal modality $[u]$ for lines are definable in the following way: $[U]A = \Box\Box A$ and $[u]\alpha = \Box\Box\alpha \wedge [\|_s]\alpha$. Seeing that $[\|_s]$ corresponds to the strong relation of parallelism, we observe that for all points X in $\mathbf{Po}$ and for all lines x in $\mathbf{Li}$:

- $\mathcal{M}, X \models \Box[\|_s]\Box A$ iff for all $Y \in \mathbf{Po}$ such that $X \neq Y$, $\mathcal{M}, Y \models A$,
- $\mathcal{M}, x \models [\|_s]\Box\Box\alpha$ iff for all $y \in \mathbf{Li}$ such that $x \neq y$, $\mathcal{M}, y \models \alpha$.

Hence, the difference modality $[D]$ for points and the difference modality $[d]$ for lines are definable in the following way: $[D]A{=}\Box[\|_s]\Box A$ and $[d]\alpha{=}[\|_s\,]\Box\Box\alpha$. To illustrate the value of our two-sorted modal language, let us remark that, in the class $\mathcal{C}_{ap}$ of all affine planes, the following formulas are valid:

Formulas of type point:	Formulas of type line:
$\Box\alpha \to \Diamond\alpha$	$\Box A \to \Diamond A$
$[U]A \to [U][U]A$	$[\|_s]\alpha \to \langle\|_s\rangle\alpha$
$[U]A \to [D]A$	$\alpha \to [\|_s]\langle\|_s\rangle\alpha$
$A \wedge [D]A \to [U]A$	$\alpha \wedge [\|_s]\alpha \to [\|_s][\|_s]\alpha$
$[U]\Box\alpha \leftrightarrow \Box\alpha \wedge \Box[\|_s]\alpha$	$[u]\alpha \to [d]\alpha$
$A \wedge \Diamond(\alpha \wedge \Diamond(\neg A \wedge [D]A)) \to \Box(\Box A \vee \alpha)$	$\alpha \wedge [d]\alpha \to [u]\alpha$
	$[u]\Box A \leftrightarrow \Box A \wedge [\|_s]\Box A$

These formulas are Sahlqvist formulas. Hence they correspond to first-order conditions on two-sorted structures. For example, the point formula $[U]A \to [U][U]A$ is related to the property of line-connectedness saying that every two points are incident with a common line whereas the point formula $[U]\Box\alpha \leftrightarrow \Box\alpha\wedge\Box[\|_s]\alpha$ and the line formula $[u]\Box A \leftrightarrow \Box A\wedge[\|_s]\Box A$ are related to the existence part of Euclid's property saying that every point not incident with a given line is incident with at least one line parallel to the given line. As for the line formula $A \wedge \Diamond(\alpha \wedge \Diamond(\neg A \wedge [D]A)) \to \Box(\Box A \vee \alpha)$, it corresponds to the normality conditions saying that every two distinct points have no more than one common incident line. Whether adding to K_2 all instances of the above formulas yields a complete axiom system for validity in $\mathcal{C}_{ag}$ is still open. However, thanks to the possibility of defining in our language the difference modalities between points or lines, we may axiomatize the validity by means of irreflexivity rules. The axiom system AFF is obtained by adding all instances of the above formulas to the basic logic together with the following special inference rules:

Irreflexivity rule of type point: "given $p \wedge [D]\neg p \to A$, prove A" where p is a proposition letter of sort point not occurring in A,

Irreflexivity rule of type line: "given $\pi \wedge [d]\neg\pi \rightarrow \alpha$, prove α" where π is a proposition letter of sort line not occurring in α,

The completeness of AFF with respect to $\mathcal{C}_{ap}$ is proved by transferring the analogous techniques based on irreflexivity rules known from the one-sorted case. See Balbiani and Goranko, 2002 for details.

Since it is possible to define the difference modalities between points or lines in our two-sorted modal language, this language is expressive enough to allow us to define formulas expressing Desarguesian and Pappian properties and to axiomatize the corresponding logics.

It is still open whether satisfiability of point formulas and line formulas within $\mathcal{C}_{ap}$ is decidable. Nevertheless, following the line of reasoning suggested by Venema, 1999, Balbiani and Goranko, 2002 proved that satisfiability within $\mathcal{C}_{ap}$ is NEXPTIME-hard.

Finally, we note that the two-sorted modal perspective in geometry is discussed further in van Benthem, 1996, where Henkin model for second-order logic are considered as two-sorted geometric structures, and in van Benthem, 1999, where space and time sorts are put together.

Concluding remarks: elementary geometry and spatial reasoning

We end this chapter with two brief remarks.

First, there is an obvious disparity in the influence and utility of modern mathematical logic to algebra and geometry: while the main (and quite deep) applications of logic to algebra are model-theoretic, the immediate role of logic in geometry is still mainly confined to axiomatizations of geometric theories and logical independence of geometric concepts and properties. While some recent model-theoretic developments (see Hodges, 1993) have deep applications to geometry, they are still far from being accessible enough to enter the geometer's toolbox. In this chapter we have just hinted that logic can say and do more to geometry than what it has so far.

Second, we admit that the topic of this chapter is not directly related to practical spatial reasoning. Yet, we believe that the issues and results discussed here are relevant to it, because quite often, spatial reasoning ignores many geometric attributes such as distances, angles, precise shapes, etc. Just as topology can sometimes be more appropriate than metric geometry for the reasoning tasks at hand, affine planes (with or without ordering) or even plain linear incidence spaces may turn out to be the right level of abstraction. For instance, this should be the case when a street map is used for orientation and routing in the city, or in designing a method for orientation in a maze. We thus see the practical value of the study of logical theories for geometric structures discussed here in offering a hierarchy of levels of abstraction, and providing

logical tools and techniques to suit the particular needs of the agent for spatial representation and reasoning.

Acknowledgments

We are indebted to Victor Pambuccian for many valuable remarks, suggestions, and references. Valentin Goranko and Ruaan Kellerman acknowledge the financial support, by means of research grants and bursaries, from the National Research Foundation and the Department of Labour of South Africa. Philippe Balbiani and Dimiter Vakarelov were supported by the ECO-NET project 08111TL. Dimiter Vakarelov also acknowledges support from the Bulgarian Ministry of Science and Education (project NIP-1510).

References

Aiello, M. and van Benthem, J. (2002). A modal walk through space. *Journal of Applied Non-Classical Logics*, 12:319–363.

Artin, E. (1957). *Geometric Algebra*. Interscience, New York.

Balbiani, P. (1998). The modal multilogic of geometry. *Journal of Applied Non-classical Logics*, 8:259–281.

Balbiani, P., Fariñas del Cerro, L., Tinchev, T., and Vakarelov, D. (1997). Modal logics for incidence geometries. *Journal of Logic and Computation*, 7(1): 59–78.

Balbiani, P. and Goranko, V. (2002). Modal logics for parallelism, orthogonality, and affine geometries. *Journal of Applied Non-Classical Logics*, 12:365–397.

Basu, S. (1999). New results on quantifier elimination over real closed fields and applications to constraint databases. *JACM*, 46(4):537–555.

Basu, S., Pollack, R., and Roy, M.-F. (1996). On the combinatorial and algebraic complexity of quantifier elimination. *Journal ACM*, 43(6):1002–1045.

Batten, L. M. (1986). *Combinatorics of Finite Geometries*. CUP.

Behnke, H., Bachmann, F., Fladt, K., and Kunle, H., editors (1974). *Fundamentals of Mathematics, Vol. II: Geometry*. MIT Press, Cambridge, Massachusetts.

Bennett, M. K. (1995). *Affine and Projective Geometry*. J. Wiley, NY.

Beth, E. and Tarski, A. (1956). Equilaterality as the only primitive notion of Euclidean geometry. *Indag. Math.*, 18:462–467.

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. CUP.

Bledsoe, W.W. and Loveland, D.W., editors (1984). *Contemporary Mathematics: Automated Theorem Proving - After 25 Years*, Providence, RI. American Mathematical Society.

Blumenthal, L. (1961). *A Modern View of Geometry*. W.H. Freeman and Company, San Francisco.

Blumenthal, L. M. and Menger, K. (1970). *Studies in Geometry*. W. H. Freeman and Company, San Francisco.

Buchberger, B. (1985). Gröbner bases: an algorithmic method in polynomial ideal theory. In Bose, N., editor, *Recent Trends in Multidimensional Systems Theory*, pages 184–232. Reidel, Dordrecht.

Buchberger, B., Collins, G. E., and Kutzler, B. (1988). Algebraic methods for geometric reasoning. Annual Review of Computer Science.

Carnap, R. (1947). *Meaning and Necessity*. University of Chicago Press.

Caviness, B. F. and Johnson, J. R., editors (1998). *Quantifier Elimination and Cylindrical Algebraic Decomposition*. Springer, New York.

Chang, C. C. and Keisler, H. J. (1973). *Model Theory*. North Holland Publishing Company, Amsterdam. 3rd ed. 1990.

Chou, S. C. (1984). Proving elementary geometry theorems using Wu's algorithm. In Bledsoe and Loveland, 1984, pages 243–286.

Chou, S. C. (1987). A method for mechanical derivation of formulas in elementary geometry. *Journal of Automated Reasoning*, 3:291–299.

Chou, S. C. (1988). An introduction to Wu's method for mechanical theorem proving in geometry. *Journal of Automated Reasoning*, 4:237–267.

Chou, S. C. (1990). Automated reasoning in geometries using the characteristic set method and Gröbner basis method. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation ISSAC'90*, pages 255–260. ACM Press.

Chou, S. C. and Gao, X. S. (1990). Ritt-Wu's decomposition algorithm and geometry theorem proving. In Stickel, M. E., editor, *Proc. CADE-10*, volume 449 of *LNCS*. Springer-Verlag.

Cohn, A. and Hazarika, S. (2001). Qualitative spatial representation and reasoning: an overview. *Fundamenta Informaticae*, 46:1–29.

Collins, G. E. (1975). Quantifier elimination for the elementary theory of real closed fields by cylindrical algebraic decomposition. *Lect. Notes Comput. Sci.*, 33:134–183.

Collins, G. E. (1998). Quantifier elimination by cylindrical algebraic decomposition—twenty years of progress. In Caviness, B. F. and Johnson, J. R., editors, *Quantifier Elimination and Cylindrical Algebraic Decomposition*, pages 8–23. Springer-Verlag, New York.

Coppel, W. A. (1998). *Foundations of Convex Geometry*. CUP.

Coxeter, H. (1969). *Introduction to Geometry*. John Wiley & Sons, NY.

Davenport, James H. and Heintz, Joos (1988). Real quantifier elimination is doubly exponential. *J. Symb. Comput.*, 5(1/2):29–35.

de Rijke, M. (1992). The modal logic of inequality. *Journal of Symbolic Logic*, 57:566–584.

de Rijke, M. (1995). The logic of Peirce algebras. *Journal of Logic, Language and Information*, 4:227–250.

Demri, S. (1996). A simple tableau system for the logic of elsewhere. In Miglioli, P., Moscato, U., Mundici, D., and Ornaghi, M., editors, *Theorem Proving with Analytic Tableaux and Related Methods*, volume 1071 of *LNAI*. Springer.

Doets, K. (1996). *Basic Model Theory*. CSLI Publications, Stanford.

Dolzmann, A., Sturm, A., and Weispfenning, V. (1998). A new approach for automatic theorem proving in real geometry. *Journal of Automated Reasoning*, 21:357–380.

Enderton, H. (1972). *A Mathematical Introduction to Logic*. Harcourt Academic Press, New York. 2nd ed. 2001.

Esser, M. (1951). Self-dual postulates for n-dimensional geometry. *Duke Math. Journal*, 18:475–479.

Esser, M. (1973). Self-dual axioms for many-dimensional projective geometry. *Trans. Amer. Math. Soc.*, 177:221–236.

Eves, H. (1972). *A Survey of Geometry*. Allyn and Bacon Inc., Boston.

Gabbay, D. (1981). An irreflexivity lemma with applications to axiomatizations of conditions in tense frames. In Monnich, U., editor, *Aspects of Philosophical Logic*, pages 67–89. Reidel, Dordrecht.

Gemignani, M. (1971). *Axiomatic Geometry*. Addison-Wesley Publ. Co.

Goldblatt, R. (1987). *Orthogonality and Space-Time Geometry*. Springer-Verlag.

Goranko, V. (1990). Modal definability in enriched languages. *Notre Dame Journal of Formal Logic*, 31:81–105.

Goranko, V. and Passy, S. (1992). Using the universal modality: gains and questions. *Journal of Logic and Computation*, 2:5–30.

Hartshorne, R. (1967). *Foundations of Projective Geometry*. W. A. Benjamin Inc., New York.

Heintz, J., Roy, M., and Solerno, P. (1990). Sur la complexite´ du principe de tarski-seidenberg. *Bull. Soc. Math. France*, 118:101–126.

Henkin, L., Suppes, P., and Tarski, A., editors (1959). *The Axiomatic Method, with Special Reference to Geometry and Physics*. North-Holland Publishing Company, Amsterdam.

Heyting, A. (1963). *Axiomatic Projective Geometry*. P. Noordhoff (Groningen) and North-Holland Publishing Company (Amsterdam).

Hilbert, D. (1950). *Foundations of Geometry*. La Salle, Illinois.

Hodges, W. (1993). *Model Theory*. Cambridge University Press.

Hughes, D. R. and Piper, F. C. (1973). *Projective Planes*. Graduate Texts in Mathematics no. 6. Springer-Verlag, New York.

Hughes, G. and Cresswell, M. (1996). *A New Introduction to Modal Logic*. Routledge.

Kapur, D. (1986). Geometry theorem proving using Hilbert's Nullstellensatz. In *Proc. of SYMSAC'86*, pages 202–208, Waterloo.

Karzel, H., Sörensen, K., and Windelberg, D. (1973). *Einführung in die Geometrie*. Studia mathematica/Mathematische Lehrbücher, Taschenbuch 1. Uni-Taschenbücher, No. 184. Vandenhoeck & Ruprecht, Göttingen.

Kordos, M. (1982). Bisorted projective geometry. *Bull. Acad. Polon. Sci. Sér. Sci. Math.*, 30(no. 9–10):429–432 (1983).

Kramer, R. (1993). The undefinability of intersection from perpendicularity in the three-dimensional Euclidean geometry of lines. *Geometriae Dedicata*, 46:207–210.

Lemon, O. and Pratt, I. (1998). On the incompleteness of modal logics of space: advancing complete modal logics of place. In Kracht, M., de Rijke, M., Wansing, H., and Zakharyaschev, M., editors, *Advances in Modal Logic: Volume 1*, pages 115–132. CSLI Publications.

Lenz, H. (1954). Zur Begründung der analytischen Geometrie. *S.-B. Math.-Nat. Kl. Bayer. Akad. Wiss. 1954*, pages 17–72 (1955).

Lenz, H. (1989). Zur Begründung der affinen Geometrie des Raumes. *Mitt. Math. Ges. Hamburg*, 11(6):763–775.

Lenz, H. (1992). Konvexität in Anordnungsräumen. *Abh. Math. Sem. Univ. Hamburg*, 62:255–285.

Lombard, M. and Vesley, R. (1998). A common axiom set for classical and intuitionistic plane geometry. *Annals of Pure and Applied Logic*, 95: 229–255.

Marx, M. (1996). Dynamic arrow logic. In Marx, M., Pólos, L., and Masuch, M., editors, *Arrow Logic and Multi-Modal Logic*, pages 109–123. CSLI Publications.

Menger, K. (1948). Independent self-dual postulates in projective geometry. *Rep. Math. Colloquium (2)*, 8:81–87.

Menger, K. (1950). The projective space. *Duke Math. Journal*, 17:1–14.

Menghini, M. (1991). On configurational propositions. *Pure Math. Appl. Ser. A.*, 2(1–2):87–126.

Meserve, B. (1983). *Fundamental Concepts of Geometry*. Dover Publ., New York, second edition.

Mihalek, R. (1972). *Projective Geometry and Algebraic Structures*. Academic Press, New York.

Pambuccian, V. (1989). Simple axiom systems for Euclidean geometry. *Mathematical Chronicle*, 18:63–74.

Pambuccian, V. (1995). Ternary operations as primitive notions for constructive plane geometry VI. *Math. Logic Quarterly*, 41:384–394.

Pambuccian, V. (2001a). Constructive axiomatizations of plane absolute, Euclidean and hyperbolic geometry. *Mathematical Logic Quarterly*, 47: 129–136.

Pambuccian, V. (2001b). Fragments of Euclidean and hyperbolic geometry. *Scientiae Mathematicae Japonicae*, 53(2):361–400.

Pambuccian, V. (2003). Sphere tangency as single primitive notion for hyperbolic and Euclidean geometry. *Forum Mathematicum*, 15:943–947.

Pambuccian, V. (2004). Axiomatizations of Euclidean geometry in terms of points, equilateral triangles or squares, and incidence. *Indagationes Mathematicae*, 15(3):413–417.

Pambuccian, V. (2006). Axiomatizations of hyperbolic and absolute geometries. In Prékopa, A. and Molnár, E., editors, *Non-Euclidean Geometries: János Bolyai Memorial Volume*, pages 119–153. Springer Verlag, New York.

Pasch, M. (1882). *Vorlesungen über neuere Geometrie*. B.G.Teubner, Leipzig.

Pieri, M. (1908). La geometria elementare istituita sulle nozioni "punto" e "sfera". *Memorie di Matematica e di Fisica della Societa Italiana delle Scienze*, 15:345–450.

Renegar, J. (1992). On the computational complexity and geometry of the first-order theory of the reals. *J. Symb. Comput.*, 13(3):255–352.

Schwabhäuser, W. and Szczerba, L. (1975). Relations on lines as primitive notions for Euclidean geometry. *Fund. Math.*, LXXXII:347–355.

Schwabhäuser, W., Szmielew, W., and Tarski, A. (1983). *Metamathematische Methoden in der Geometrie*. Springer-Verlag, Berlin.

Scott, D. (1956). A symmetric primitive of Euclidean geometry. *Indag. Math.*, 18:456–461.

Scott, D. (1959). Dimension in elementary Euclidean geometry. In Henkin et al., 1959, pages 53–67.

Segerberg, K. (1981). A note on the logic of elsewhere. *Theoria*, 47:183–187.

Seidenberg, A. (1954). A new decision method for elementary algebra. *Ann. Math.*, 60:365–374.

Spaan, E. (1993). *Complexity of Modal Logics*. PhD thesis, University of Amsterdam.

Stebletsova, V. (2000). *Algebras, Relations, Geometries*. PhD thesis, Zeno Institute of Philosophy, Univ. of Utrecht.

Stockmeyer, L. J. (1977). The polynomial-time hierarchy. *Theoretical Computer Science*, 3(1):1–22.

Szczerba, L. (1972). Weak general affine geometry. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.*, 20:753–761.

Szczerba, L. and Tarski, A. (1965). Metamathematical properties of some affine geometries. In Bar-Hillel, Y., editor, *Proceedings of the 1964 International Congress for Logic, Methodology and Philosophy of Science*, Studies in Logic and the Foundations of Mathematics, Amsterdam. North-Holland Publishing Company.

Szczerba, L. and Tarski, A. (1979). Metamathematical discussion of some affine geometries. *Fund. Math.*, 104:155–192.

Szmielew, W. (1959). Some metamathematical problems concerning elementary hyperbolic geometry. In Henkin et al., 1959, pages 30–52.

Szmielew, W. (1983). *From Affine to Euclidean Geometry: An Axiomatic Approach*. D. Reidel and PWN-Warsaw.

Tarski, A. (1949a). On essential undecidability. *Journal of Symbolic Logic*, 14:75–76.

Tarski, A. (1949b). Undecidability of the theories of lattices and projective geometries. *Journal of Symbolic Logic*, 14:77–78.

Tarski, A. (1951). A decision method for elementary algebra and geometry. Technical report, UCLA. Prepared for publ. by J. McKinsey.

Tarski, A. (1956). A general theorem concerning primitive notions of Euclidean geometry. *Indag. Math.*, 18:468–474.

Tarski, A. (1959). What is elementary geometry? In Henkin et al., 1959, pages 16–29.

Tarski, A. (1967). The completeness of elementary algebra and geometry. Technical report, Institut Blaise Pascal, Paris.

Tarski, A. and Givant, S. (1999). Tarski's system of geometry. *Bull. of Symb. Logic*, 5(2):175–214.

Tarski, A. and Mostowski, A. (1949). Undecidability in the arithmetic of integers and in the theory of rings. *Journal of Symbolic Logic*, 14:76.

Tarski, A., Mostowski, A., and Robinson, R. (1953). *Undecidable theories*. North-Holland.

van Benthem, J. (1983). *The Logic of Time*. Kluwer.

van Benthem, J. (1984). Correspondence theory. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic: Volume II*, pages 167–247. Reidel.

van Benthem, J. (1994). A note on dynamic arrow logics. In van Eijck, J. and Visser, A., editors, *Logic and Information Flow*, pages 15–29. MIT Press.

van Benthem, J. (1996). Complexity of contents versus complexity of wrappings. In Marx, M., Masuch, M., and Pólos, editors, *Arrow Logic and Multimodal Logic*, pages 203–219. CSLI Publications, Stanford.

van Benthem, J. (1999). Temporal patterns and modal structure. *Logic Journal of IGPL*, 7:7–26.

Veblen, O. (1904). A system of axioms for geometry. *Transactions of the American Mathematical Society*, 5:343–384.

Veblen, O. (1914). The foundations of geometry. In Young, J., editor, *Monographs on topics of modern mathematics, relevant to the elementary field*, pages 1–51. Longsman, Green, and Company, New York.

Venema, Y. (1993). Derivation rules as anti-axioms in modal logic. *Journal of Symbolic Logic*, 58:1003–1034.

Venema, Y. (1999). Points, lines and diamonds: a two-sorted modal logic for projective planes. *Journal of Logic and Computation*, 9:601–621.

von Plato, J. (1995). The axioms of constructive geometry. *Annals of Pure and Applied Logic*, 76:169–200.

von Wright, G. (1979). A modal logic of place. In Sosa, E., editor, *The Philosophy of Nicholas Rescher*, pages 65–73. Reidel.

Wu, W. (1984). Some recent advances in mechanical theorem proving in geometries. In Bledsoe and Loveland, 1984, pages 235–242.

Wu, W. (1986). Basic principles of mechanical theorem proving in geometries. *Journal of Automated Reasoning*, 2(4):221–252.

Ziegler, M. (1982). Einige unentscheidbare Körpertheorien. *Enseign. Math. (2)*, 28(3–4):269–280.

Chapter 8

LOCALES AND TOPOSES AS SPACES

Steven Vickers
University of Birmingham

Second Reader

Guram Bezhanishvili
New Mexico State University

1. Introduction

Mac Lane and Moerdijk, 1992, in their thorough introduction to topos theory, start their Prologue by saying:

> A startling aspect of topos theory is that it unifies two seemingly wholly distinct mathematical subjects: on the one hand, topology and algebraic geometry, and on the other hand, logic and set theory. Indeed, a topos can be considered both as a "generalized space" and as a "generalized universe of sets".

This dual nature of topos theory is of great importance, and one can quite reasonably understand Grothendieck's name "topos" as meaning "that of which topology is the study". Mac Lane and Moerdijk are unquestionably masters of the spatial nature of toposes, yet one could easily read through their book without grasping it. The mathematical technology is so firmly expressed in the set theory and the logic that the spatiality is obscured.

The aim in this chapter is to provide a reader's guide to the spatial content of the major texts. Those texts can also provide a more detailed account of original sources and other applications than has been possible here.

We have on the one hand, the logic and set theory, and, on the other, the topology. In a nutshell, the topos connection between them is that the topos acts like a "Lindenbaum algebra" (of formulae modulo equivalence) for a logical theory whose models are the points of a space.

The prototype is Stone's Representation Theorem for Boolean algebras, which relates propositional logic to Hausdorff, totally disconnected topology.

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 429–496.
© 2007 *Springer.*

However, it takes some work to develop the idea to its full generality. First, the logic is not at all ordinary classical logic. It is an infinitary positive logic known as *geometric* logic. Second, we are in general talking about predicate theories, and for these the appropriate notion of Lindenbaum algebra is not straightforward. It is really the "category of sets generated by the theory". Lastly, "space" of points is not an ordinary topological space—it is a real generalization.

However, the propositional fragment of the predicate logic does correspond more or less to ordinary topological spaces. As a rough picture of the correspondence, in the propositional case we find:

- space $\sim$ logical theory
- point $\sim$ model of the theory
- open set $\sim$ propositional formula
- sheaf $\sim$ predicate formula
- continuous map $\sim$ transformation of models that is definable within geometric logic

These "propositional toposes" are called *localic*, or (with slight abuse of language) *locales*. They are equivalent to the locales introduced in—say—Johnstone, 1982 or Vickers, 1989.

Now the topos theorists discovered some deep facts about the interaction between continuous maps and the logic and set theory of toposes. A map $f : X \to Y$ gives a geometric morphism between the corresponding toposes of sheaves. A topos is sufficiently like the category of sets that a kind of set theory can be modelled in it. Roughly speaking, in sheaves over X it is set theory "continuously parametrized by a variable point of X". The map f then comes to be seen as a "generalized" point of Y, parametrized by a point of X, and this is a point of Y in the non-standard set theory of sheaves over X. So by allowing topological reasoning to take place in toposes instead of in the category of ordinary sets, one gains a simple way to reason about the generalized points of Y—in other words the maps into Y.

However, to make this trick work one has to reason constructively because the internal logic of toposes is not in general classical. And constructive topology does not work well unless one replaces topological spaces with locales. For instance, the Tychonoff theorem and the Heine-Borel theorem hold constructively for locales but not for topological spaces.

Now there is a well known drawback to locales. They do not in general have enough points and for this reason are normally treated with an opaque "point-free" style of argument. However, they do have enough *generalized* points. Since constructive reasoning gives easy access to these, it also allows locales to

be discussed in a spatial way in terms of their points. We in fact get a cohesive package of mathematical deals.

1 Constructive reasoning allows maps to be treated as generalized points.

2 Locales give a better constructive topology (better results hold) than ordinary spaces.

3 The constructive reasoning makes it possible to deal with locales as though they were spaces of points.

What's more, the more stringent *geometric* constructivism has an intrinsic continuity—one might almost say it is the logical essence of continuity. The effect of this is that constructions described in conformity with its disciplines are automatically continuous.

The prime aim of this chapter is to explain how this deep connection between logic and topology works out. However, as a spinoff we find that "generalized spaces" corresponding to toposes become more accessible. They are spaces in which the opens are insufficient to define the topological structure, and sheaves have to be used instead.

These ideas are not essentially new. They have been a hidden part of topos theory from the start. Some writers, such as Wraith, 1979, have made quite explicit use of the virtues of geometric logic. Our aim here is to make them less hidden. At the same time we shall also stress a peculiarity of geometric logic, namely that it embodies a geometric *type theory*. This provides a more naturally mathematical mode of working in geometric logic.

For further reading as a standard text on topos theory, we particularly recommend Mac Lane and Moerdijk, 1992. The standard reference text (Johnstone, 2002a, Johnstone, 2002b) is much more complete and ultimately indispensable. In particular, it treats in some depth the notion of "geometric type construct" that is very important for us. However, it can be impenetrable for beginners.

Though the chapter is so closely linked to toposes, many of its techniques can also be used in other (and distinct) constructive foundations such as formal topology in predicative type theory (Sambin, 1987). We refer to Vickers, 2006 and Vickers, 2005 for some of the connections.

2. Opens as propositions

Since Tarski, 1938 it has been known that topologies—by which we mean specifically the lattices of open sets for topological spaces—can provide models for intuitionistic propositional logic.

For discrete topologies, i.e. powersets, this is no surprise. Classical propositional logic can be embedded in classical predicate logic by translating each proposition into a predicate with a single variable x, and then the standard semantics interprets each proposition as a subset of the carrier for x. Logical

connectives translate directly into the corresponding set-theoretic operations in the powerset, and classicality of the logic corresponds to the fact (in classical mathematics) that powersets are Boolean algebras.

What is interesting is that when we replace the powerset by a topology (on X, say), there is still enough lattice structure to model intuitionistic logic. The connectives $\wedge$ and $\vee$ can still be translated to $\cap$ and $\cup$, which both preserve openness. However the direct set-theoretic correspondent with negation is complementation, and that does not preserve openness. If proposition P is interpreted as open set U, then $\neg P$ is interpreted as the *interior* of the complement $X - U$. Similarly, if also Q is interpreted as V, then $P \to Q$ is interpreted as the interior of $(X - U) \cup V$.

These latter operations can both be defined directly in terms of the complete lattice structure of the topology. The interior of $X - U$ is the join of all those opens W such that $W \cap U = \emptyset$, while the interior of $(X - U) \cup V$ is the join of those W such that $W \cap U \subseteq V$. Every topology is a Heyting algebra; and since intuitionistic logic freely expresses Heyting algebra structure, any interpretation of the propositional symbols can be extended to all formulae in a way that respects intuitionistic equivalence.

DEFINITION 8.1 *A* frame *is a complete lattice in which the following* frame distributivity law *holds ($a \in A$, $S \subseteq A$):*

$$a \wedge \bigvee S = \bigvee \{a \wedge b \mid b \in S\}.$$

A frame homomorphism *is a function between frames that preserves joins and finite meets.*

Every topology is a frame, because the axioms for opens sets tell us that $\wedge$ and $\bigvee$ are set theoretic $\cap$ and $\bigcup$.

PROPOSITION 8.2 *Every frame A is a Heyting algebra.*

Proof We must show that for every $a, b \in A$ there is an element $a \to b \in A$ such that for all $c \in A$,

$$(*) \qquad c \leq a \to b \Leftrightarrow c \wedge a \leq b.$$

(If this element exists, then it is unique.) We can define

$$a \to b = \bigvee \{x \mid x \wedge a \leq b\}.$$

Now the ($\Leftarrow$) direction in (*) is immediate, while the ($\Rightarrow$) follows from frame distributivity, which implies that $(a \to b) \wedge a \leq b$. QED

In fact, frames and complete Heyting algebras are the same things. However, we distinguish between the notions because of the homomorphisms. A frame

homomorphism does not necessarily preserve the Heyting arrow, and so is not necessarily a Heyting algebra homomorphism.

Tarski, 1938 went further, and showed a stronger property if the space X is a dense-in-itself separable metric space (such as the real line). If P_i ($i \in \mathbb{N}$) is a countable family of propositional symbols, then an intuitionistic formula ϕ in the P_is is an intuitionistic theorem iff, for every interpretation of the P_is in ΩX, the corresponding interpretation of ϕ is the whole of X. This was explained further in McKinsey and Tarski, 1944 using an embedding of intuitionistic propositional logic in the classical modal logic **S4**, which could then be interpreted in the powerset of X with the $\Box$ modality corresponding to the interior operator—for further details, see 5.

2.1 Lindenbaum algebras for classical logic

A Lindenbaum algebra is a lattice of formulae modulo provable equivalence, and they and their generalizations will be key to the whole of this chapter. Let us review how it works for classical logic, where the connection with topology is essentially Stone's Representation Theorem for Boolean algebras (see Johnstone, 1982). If Σ is a propositional signature (i.e. a set of propositional symbols) then we write Sen_Σ for the set of sentences constructed over Σ using a classically adequate set of connectives. If T is a theory over Σ (a set of sentences), then an equivalence relation $\equiv_T$ can be defined by

$$\phi \equiv_T \psi \text{ iff } T \vdash (\phi \leftrightarrow \psi) \text{ in classical logic}$$

and the Lindenbaum algebra for T is defined as $\mathcal{LA}(\Sigma, T) = \mathsf{Sen}_\Sigma / \equiv_T$. It is a Boolean algebra.

The central idea is to use the dual nature of homomorphisms $g : \mathcal{LA}(\Sigma, T) \to \mathcal{LA}(\Sigma', T')$.

Logically, g is a logical translation of (Σ, T) into (Σ', T'). It translates propositional symbols into sentences (modulo equivalence) and preserves theoremhood. Isomorphism gives a natural presentation-independent notion of equivalence of theories, by mutual translatability.

Spatially, g provides a transformation of models, from (Σ', T') to (Σ, T) (note the reversal of direction). It is this spatial view that provides the link with topology, for a model transformation arises this way iff it is continuous with respect to certain topologies on the model spaces. This is shown by Stone's Theorem.

Because classical logic is complete, the transformation of ordinary models suffices to determine the Lindenbaum algebra homomorphism. However, we also get an alternative view by considering generalized models. Then there is a *generic* model whose transformation determines that of all the other (specific) models, and this idea becomes important for incomplete logics.

Let A be a Boolean algebra. An *interpretation* of Σ in A is a function $M : \Sigma \to A$, and this extends uniquely to a function $\overline{M} : \mathsf{Sen}_\Sigma \to A$ that evaluates the connectives by the corresponding Boolean operations on A. Then M is a *model* of T iff $\overline{M}(\phi) = 1$ for every $\phi \in T$. Models are preserved by Boolean algebra homomorphisms $f : A \to B$ – the composite $f \circ M$ is also a model in B. We write $\mathrm{Mod}_A(T)$ for the set of models of T in A.

Those were the generalized models referred to above. The standard models, interpreting propositions as truth values, are found by taking $A = \mathbf{2} = \{0, 1\}$.

The *generic* model M_T of T is a particular model in the Lindenbaum algebra. It interprets each proposition symbol $P \in \Sigma$ as the equivalence class of P as sentence. Clearly, by definition of the Lindenbaum algebra we have $\phi \equiv_T \top$ for every $\phi \in T$, and so $\overline{M_T}(\phi) = 1$, so M_T is a model. It has a universal property: any model can be got, uniquely, by applying a Boolean algebra homomorphism to the generic model.

PROPOSITION 8.3 *Let (Σ, T) be a classical propositional theory, and A a Boolean algebra. Then the function $f \mapsto f \circ M_T$, taking Boolean algebra homomorphisms $\mathcal{LA}(\Sigma, T) \to A$ into $\mathrm{Mod}_A(T)$, is a bijection.*

Proof Let $M : \Sigma \to A$ be a model of T. Suppose $\phi \equiv_T \psi$. The classical proof of $T \vdash (\phi \leftrightarrow \psi)$ will involve only finitely many elements of T, say $t_1, \ldots, t_n$. Because of the nature of classical proofs (this requires some checking) it will imply that $\overline{M}(t_1) \wedge \cdots \wedge \overline{M}(t_n) \leq \overline{M}(\phi) \leftrightarrow \overline{M}(\psi)$ in A. But each $\overline{M}(t_i)$ is 1, because M is a model, and hence $\overline{M}(\phi) \leftrightarrow \overline{M}(\psi) = 1$ and $\overline{M}(\phi) = \overline{M}(\psi)$. It follows that $\overline{M}$ factors (uniquely) through $\mathcal{LA}(\Sigma, T)$ as $f \circ M_T$ where f is a Boolean algebra homomorphism. QED

The generic model M_T corresponds to the identity homomorphism on $\mathcal{LA}$ (Σ, T).

As a consequence of the proposition, we see there is a function

$$\vDash : \mathrm{Mod}_A(T) \times \mathcal{LA}(\Sigma, T) \to A$$

with $\vDash (M, \phi)$ the image of ϕ under the homomorphism corresponding to M. In the particular case of $A = \mathbf{2}$, $\vDash (M, \phi) = 1$ iff $M \vDash \phi$ in the usual sense.

Now consider a homomorphism $g : \mathcal{LA}(\Sigma, T) \to \mathcal{LA}(\Sigma', T')$. By the proposition, g corresponds to a model of T in $\mathcal{LA}(\Sigma', T')$. This gives a logical translation of (Σ, T) in (Σ', T') – the propositional symbols in Σ are interpreted as formulae over Σ', in such a way that the axioms in T all become provable from T'.

But one can also view this from the model side. A model of T' in A is a homomorphism from $\mathcal{LA}(\Sigma', T')$ to A, and precomposing this with g gives a homomorphism from $\mathcal{LA}(\Sigma, T)$ to A, in other words a model of T. (Note the reversal of direction!) Thus g gives a uniform way of transforming models of T' into models of T.

Trivially, $g = g \circ \mathrm{Id}_{\mathcal{LA}(\Sigma',T')}$ is completely determined by its transformation of the generic model. The generic model is non-standard, but we also find g is determined by its transformation of the standard models.

PROPOSITION 8.4 *Let (Σ, T) and (Σ', T') be two propositional theories, and let $f, g : \mathcal{LA}(\Sigma, T) \to \mathcal{LA}(\Sigma', T')$ be two homomorphisms inducing model transformations $F, G : \mathrm{Mod}_{\mathbf{2}}(T') \to \mathrm{Mod}_{\mathbf{2}}(T)$. If $F = G$ then $f = g$.*

Proof It suffices to show that $f(P) \equiv_{T'} g(P)$ for every $P \in \Sigma$. By completeness, it suffices to show that for every standard model M' of T', $f(P)$ and $g(P)$ have the same truth value at M'. But those truth values are the same as those for P at $F(M')$ and $G(M')$ respectively, and they are equal. QED

Not every transformation of standard models is induced by a homomorphism. Stone's Representation Theorem represents $\mathcal{LA}(\Sigma, T)$ using a topological space $\mathrm{Mod}_{\mathbf{2}}(T)$, with sets of the form $\{M \mid M \vDash \phi\}$ ($\phi \in \mathcal{LA}(\Sigma, T)$) providing a base of opens. (In fact, they are the clopens.) This is a *Stone space* – Hausdorff and totally disconnected (see Johnstone, 1982). The Theorem shows that the homomorphisms correspond to the *continuous* maps between the model spaces.

DEFINITION 8.5 *Let (Σ, T) and (Σ', T') be two propositional theories. We define a* map *from (Σ, T) to (Σ', T') to be a homomorphism from $\mathcal{LA}(\Sigma', T') \to \mathcal{LA}(\Sigma, T)$ (or, equivalently, a model of (Σ', T') in $\mathcal{LA}(\Sigma, T)$).*

Note – In this chapter, the word "map" will always carry connotations of continuity. A map between topological spaces is understood to be continuous.

By emphasizing the model transformations, and defining "maps" to go in the same direction (opposite to that of the homomorphisms), we try to foster a view that the theory represents its "space of models". Propositional theories and maps between them form a category. Here are some simple but important examples.

EXAMPLE 8.6 *(1) The theory $(\emptyset, \emptyset)$ (no symbols or axioms) corresponds to the one-element space 1. It has a unique, vacuous model in any Boolean algebra and is final in the category of theories. $\mathcal{LA}(\emptyset, \emptyset) = \mathbf{2}$, and for any theory (Σ, T) the maps from $(\emptyset, \emptyset)$ to (Σ, T) are equivalent to the standard models of (Σ, T).*

(2) The inconsistent theory $(\emptyset, \{\bot\})$ corresponds to the empty space $\emptyset$. It has no model except in the one-element Boolean algebra $\mathbf{1}$, which is $\mathcal{LA}(\emptyset, \{\bot\})$. It is initial in the category of theories.

(3) The theory $(\{P\}, \emptyset)$ corresponds to the discrete 2-element space 2, having two standard models $P \mapsto 0$ and $P \mapsto 1$. Models in A are elements of A. In Stone's Theorem, this corresponds to the fact that for any space X, clopens

are equivalent to maps $X \to 2$. Its Lindenbaum algebra $\mathcal{LA}(\{P\}, \emptyset)$ is the Boolean algebra $\mathbf{4} = \{0, P, \neg P, 1\}$, which is freely generated by P.

2.2 Frames as Lindenbaum algebras

The Stone topologies, with the Boolean algebra of clopens forming a base and corresponding to classical propositional logic, are very special. We generalize this by changing to propositional *geometric* logic, for which the Lindenbaum algebras are *frames*, playing the role of topologies (with all opens, not just clopens).

From Tarski's results on interpreting intuitionistic logic in topologies, one might expect the logic here to be intuitionistic. However, full intuitionistic logic is too strong. If f is a continuous map, then its inverse image function, restricted to open sets, will act as the corresponding homomorphism between Lindenbaum algebras. In general this does not preserve the Heyting arrow, though it is a frame homomorphism. Hence we need a logic that corresponds to the structure of frame rather than of Heyting algebra.

DEFINITION 8.7 *Let Σ be a set (of propositional symbols).* Geometric formulae *over Σ are constructed from the symbols in Σ using $\top$ (true), $\wedge$ and arbitrary – possibly infinitary – disjunctions $\bigvee$. A* geometric theory *over Σ is a set of axioms of the form $\phi \to \psi$, where ϕ and ψ are geometric formulae.*

Coherent *formulae and theories are defined in the same way, but without any infinitary disjunctions. This is sometimes known as* positive *logic.*

Note that because of the limitations of the logic, a theory is not simply a set of formulae. The logical rules are best described in sequent form. (A theory is in effect a set of axiomatic sequents, and we shall often write its axioms as sequents, using $\vdash$. We shall also write $\vdash\dashv$ for bidirectional entailment.) The rules are *identity*

$$\phi \vdash \phi,$$

cut

$$\frac{\phi \vdash \psi \qquad \psi \vdash \chi}{\phi \vdash \chi},$$

the *conjunction* rules

$$\phi \vdash \top, \quad \phi \wedge \psi \vdash \phi, \quad \phi \wedge \psi \vdash \psi, \quad \frac{\phi \vdash \psi \quad \phi \vdash \chi}{\phi \vdash \psi \wedge \chi},$$

the *disjunction* rules

$$\phi \vdash \bigvee S \quad (\phi \in S), \quad \frac{\phi \vdash \psi \quad (\text{all } \phi \in S)}{\bigvee S \vdash \psi}$$

and *frame distributivity*

$$\phi \wedge \bigvee S \vdash \bigvee\{\phi \wedge \psi \mid \psi \in S\}.$$

Note that $\bigvee \emptyset$ plays the role of $\perp$ (false). Note also that frame distributivity allows us to reduce every formula to a disjunction of finite conjunctions of symbols from Σ. Hence although the formulae as defined syntactically form a proper class, modulo equivalence they form a set (classically, and according to at least some constructive foundations).

We write $\Omega[T]$ for the Lindenbaum algebra of T, i.e. the set of geometric formulae modulo equivalence provable from T. The logical rules imply that it is a frame. There is an obvious notion of model of T in any frame, and $\Omega[T]$ has a particular generic model M_T given by interpreting each propositional symbol as its equivalence class of formulae.

PROPOSITION 8.8 *(cf. Proposition 8.3.) Let (Σ, T) be a geometric theory, and let A be a frame. Then the function $f \mapsto f \circ M_T$, taking frame homomorphisms $\Omega[T] \to A$ into $\mathrm{Mod}_A(T)$, is a bijection.*

Proof If $M : \Sigma \to A$ is a model then it extends to a function $\overline{M}$ on the class of formulae. All the logical rules will be valid in A under this interpretation, and the axioms in T will all hold because M is a model, so it follows that $\overline{M}$ factors (uniquely) via $\Omega[T]$. QED

Standard models are given, as usual, by interpreting the propositional symbols as truth values. However, we write Ω for the frame of truth values, by contrast with $\mathbf{2}$ for the Boolean algebra. (This follows the topos-theoretic notation.) Constructively they are different, allowing for the fact that geometric logic is a positive logic. A geometric truth value is equivalent to a subset of a singleton, while a Boolean truth value is a *decidable* subset of a singleton.

DEFINITION 8.9 *Let A be a frame. Then the geometric theory Th_A is presented as follows. For the signature Σ, introduce a propositional symbol P_a for each $a \in A$, and then take axioms*

$$\begin{aligned} P_a &\to P_b && (a \leq b \text{ in } A) \\ P_a \wedge P_b &\to P_{a \wedge b} && (a, b \in A) \\ \top &\to P_1 \\ P_{\bigvee S} &\to \textstyle\bigvee_{a \in S} P_a && (S \subseteq A) \end{aligned}$$

All the theory is saying is that the finite meets and arbitrary joins in A should be treated logically as finite conjunctions and arbitrary disjunctions. Hence the connectives of propositional geometric logic correspond directly to the frame

structure. From this it follows that the models of Th_A in a frame B are the frame homomorphisms from A to B, and so we see that $\Omega[\mathrm{Th}_A] \cong A$.

A standard model of Th_A can also be described by saying which propositional symbols P_a are assigned the truth value $\mathbf{true}$, and hence by a subset $F \subseteq A$ satisfying certain conditions corresponding to the axioms. The first axiom says that F is an *upper* subset of A, the next two that F is a *filter*, and then the fourth that it is a *completely prime* filter. (Note: the standard texts contain various other descriptions of the standard models, but they are constructively inequivalent.)

2.3 Locales

Let us define, conceptually, a *locale* to be a "propositional geometric theory pretending to be a space", using the ideas of Sec. 2.1, which took the logical theory as the starting point. That is to say, the locale *is* the theory, but repackaged in a spatial language of points and maps instead of models and Lindenbaum algebra homomorphisms. What makes this repackaging significant is the fact that geometric logic is incomplete—in general, there are not enough standard models to account for all the frame homomorphisms (cf. Proposition 8.4). Thus the spatial side (in terms of standard models) and the logical side (in terms of Lindenbaum algebras) become mathematically inequivalent. However, the logical side still contains good topological results; indeed, in constructive mathematics they are often better than the spatial ones. The localic repackaging makes it much easier to see this topological content.

The usual definition is that a locale *is* a frame. We prefer to say it is the propositional geometric theory, and that it *has* a frame. This makes it easier to see locales as a special case of toposes, which arise from predicate geometric theories. In addition, in certain foundational schools such as predicative type theory, the frames are problematic. They are constructed using the powerset, and that is impredicative. The main account in this school is the formulation as "formal topology" (Sambin, 1987).

A formal topology gives a base S (so every frame element is to be a join of base elements) and the *cover* relation $\lhd$, which describes when one basic open is to be covered by a set of others. This then corresponds to a propositional geometric theory in which S provides the propositional symbols, and the cover relation provides axioms to say one symbol entails a disjunction of others. (For these purposes, the notion of map can be defined in a more primitive way that does not rely on constructing the frame.) The variant notion of *inductively generated formal topology* (Coquand et al., 2003) is even closer to the propositional geometric theory in that it does not require the complete cover relation but just a part from which the rest can be deduced.

Let us review the ideas of Sec. 2.1 in the light of Sec. 2.2.

DEFINITION 8.10 *1 A* locale *is (presented by) a propositional geometric theory. If the theory is T, we write $[T]$ for the locale. The locale $[T]$ should be conceptualized as "the space of models of T".*

2 If X is a locale, then ΩX denotes its Lindenbaum algebra, a frame.

3 The opens *of X are the elements of ΩX.*

4 If X and Y are locales, then a map *$f : X \to Y$ is a frame homomorphism $f^* : \Omega Y \to \Omega X$. We write $\mathrm{Map}(X, Y)$ for the set of maps from X to Y. Locales and maps form a category* **Loc***, dual to the category* **Fr** *of frames.*

5 If X and Y are locales then $\mathrm{Map}(X, Y)$ is partially ordered by the specialization order*, $f \sqsubseteq f'$ if $f^*(U) \leq f'^*(U)$ for every $U \in \Omega Y$.*

6 If X and W are locales then the (generalized) points *of X* at stage (of definition) *W are the maps $W \to X$. Think of these as points of X "continuously parametrized by a variable point of W". The points of $[T]$ are just the models of T in ΩW.*

7 If X is a locale then the identity map $X \to X$, a point of X at stage X, is the generic *point of X.*

8 If $f : X \to Y$ is a map, then postcomposition $f \circ -$ transforms points of X (at any stage) to points of Y (at the same stage). We shall often write $f(x)$ for $f \circ x$.

The specialization order is already present (as a preorder) in ordinary topology: $x \sqsubseteq x'$ if every neighbourhood of x also contains x' (i.e. x is in the closure of $\{x'\}$). It is often neglected there, because for Hausdorff spaces (more precisely for T_1 spaces) it is discrete: $x \sqsubseteq x'$ iff $x = x'$.

An important fact about the specialization order is that it has directed joins. In any poset $(P, \leq)$, a family of elements $(x_i)_{i \in I}$ is *directed* if I is inhabited and, for any $i, j \in I$, there is some $k \in I$ such that $x_i \leq x_k$ and $x_j \leq x_k$. We also say P is a *directed complete* poset (or *dcpo*) if it has a join for every directed family. We use an arrow, as in $\bigsqcup_i^{\uparrow} x_i$, to indicate that a join is of a directed family.

PROPOSITION 8.11 *Let X and Y be locales. Then $\mathrm{Map}(X, Y)$ is directed complete with respect to $\sqsubseteq$. The directed joins are preserved by composition on either side.*

Proof Let $(f_i)_{i \in I}$ be directed in $\mathrm{Map}(X, Y)$. The join $\bigsqcup_i^{\uparrow} f_i$ is given as a frame homomorphism by

$$(\bigsqcup_i^{\uparrow} f_i)^*(U) = \bigvee_{i \in I} f_i^*(U).$$

QED

The directed joins are less familiar from ordinary topology. This is partly because so many familiar spaces have discrete specialization order, but also because in the absence of sobriety (Sec. 2.4) the directed joins may be missing. However, they are fundamental in computer science and provide a means for providing the semantics of recursive algorithms. (See, e.g., Plotkin, 1981, Gierz et al., 1980, Vickers, 1989.) From the proposition we see one essential feature of maps, namely that they preserve directed joins of points. (This is known as Scott continuity.)

DEFINITION 8.12 *(1) The* one-point *locale* 1 *is presented by the empty theory over the empty signature.* $\Omega 1$ *is* Ω. *(Note—we shall also write* 1 *for a singleton* set. *In practice this ambiguity should not cause problems.) The* global *points of a locale* X *are its points at stage* 1, *i.e. the maps* $1 \to X$. *Thus the global points of* $[T]$ *are the standard models of* T.

(2) The empty *locale* $\emptyset$ *is presented by the inconsistent theory* $\{\top \to \bot\}$ *over the empty signature.* $\Omega\emptyset$ *is a one-element frame. It has no points except at stage* $\emptyset$.

(3) The Sierpiński *locale* $\mathbb{S}$ *is presented by the empty theory over a one-element signature* $\{P\}$. *Its points are equivalent to subsets of the set* 1. *We usually write* $\top$ *for the subset* 1 *itself (an open point), and* $\bot$ *for the empty subset (a closed point). Note that* $\bot \sqsubseteq \top$.

REMARK 8.13 *The opens* U' *of any locale* X *are equivalent to the maps* $U : X \to \mathbb{S}$, *with* $U' = U^*(P)$. *If* $f : X \to Y$ *is a map, then*

$$f^*(U') = f^*(U^*(P)) = (U \circ f)^*(P),$$

and hence corresponds to $U \circ f$. *Hence we can talk about opens and inverse image functions purely in the language of maps.*

2.4 Locales compared with spaces

Now we have this language of points, opens and maps, all deriving from the single notion of geometric theory, we shall compare it with ordinary topology.

Given a locale X, let us write $\mathrm{pt}(X)$ for its set of global points, maps $x : 1 \to X$. If $U : X \to \mathbb{S}$ is an open, then the composite $U \circ x$ is a global point of $\mathbb{S}$, and hence a subset of 1. We write $x \vDash U$ iff $U \circ x = \top$ (i.e. $x^*(U) = 1$), and $\mathrm{ext}(U)$ (the *extent* of U) for $\{x \in \mathrm{pt}(X) \mid x \vDash U\}$. Because $x^* : \Omega X \to \Omega$ is a frame homomorphism, we find that the sets $\mathrm{ext}(U)$ form a topology on $\mathrm{pt}(X)$.

Now let $f : X \to Y$ be a map of locales, giving a point transformer $\mathrm{pt}(f) : \mathrm{pt}(X) \to \mathrm{pt}(Y)$. If V is an open of Y, then

$$\begin{aligned} x \in \mathrm{pt}(f)^{-1}(\mathrm{ext}(V)) &\Leftrightarrow f \circ x \vDash V \\ &\Leftrightarrow V \circ f \circ x = \top \\ &\Leftrightarrow x \vDash \mathrm{ext}(f^*(V)). \end{aligned}$$

It follows that $\mathrm{pt}(f)^{-1}(\mathrm{ext}(V)) = \mathrm{ext}(f^*(V))$, and so $\mathrm{pt}(f)$ is continuous.

- For each locale X, its global points form a topological space $\mathrm{pt}(X)$.
- For each map of locales, the corresponding transformation $\mathrm{pt}(f)$ of global points is continuous.

This looks promising, but there is not an exact match between locales and topological spaces, and we need to understand that. The central connection, a categorical adjunction, is summarized in the following result. For an element x of a topological space X, we write $\mathfrak{N}_x = \{U \in \Omega X \mid x \in U\}$ for the set of open neighbourhoods of x. This is a completely prime filter in ΩX .

PROPOSITION 8.14 *Let X be a topological space and Y a locale. Then there is a bijection between*

1. *maps (continuous, as always) $f : X \to \mathrm{pt}(Y)$, and*
2. *maps $g : [\mathrm{Th}_{\Omega X}] \to Y$ (homomorphisms $g^* : \Omega Y \to \Omega X$).*

Proof Consider the following condition on pairs (f, ϕ) where $f : X \to \mathrm{pt}(Y)$ and $\phi : \Omega Y \to \Omega X$ are arbitrary functions:

$$(\forall x \in X,\ \forall V \in \Omega Y)\ (f(x) \in \mathrm{ext}(V) \leftrightarrow x \in \phi(V)).$$

This is equivalent to

$$(\forall V \in \Omega Y)\ \phi(V) = f^{-1}(\mathrm{ext}(V))$$

and, considering the points of $\mathrm{pt}(Y)$ as completely prime filters of ΩX and remembering that $f(x) \in \mathrm{ext}(V)$ iff $V \in f(x)$, to

$$(\forall x \in X)\ f(x) = \phi^{-1}(\mathfrak{N}_x).$$

It follows that ϕ is determined by f, and f is determined by ϕ.

Under these conditions, it follows that f is continuous and ϕ is a frame homomorphism. Conversely, if f is continuous then inverse image f^{-1} gives a corresponding ϕ; and given a frame homomorphism ϕ, we find that each $\phi^{-1}(\mathfrak{N}_x)$ is a completely prime filter. QED

The first mismatch between spaces and locales is that *not every space comes from a locale.*

Let $(X, \Omega X)$ be a topological space, with ΩX the topology—the family of open sets. Consider the locale $[\mathrm{Th}_{\Omega X}]$, whose global points are the completely prime filters of ΩX. For every point $x \in X$, its open neighbourhood filter $\mathfrak{N}_x$ is a completely prime filter. However, two points x and y might have the same open neighbourhood filter—every open containing x also contains y, and vice versa. A space is called T_0 if this never happens, i.e. if $\mathfrak{N}_x = \mathfrak{N}_y$ then $x = y$.

In addition, there may be a completely prime filter that is not $\mathfrak{N}_x$ for any point x.

EXAMPLE 8.15 *Consider the set $\mathbb{N}$ of natural numbers with a topology in which the opens are the upper sets. The completely prime filter of non-empty upper sets is not the open neighbourhood filter of any point.*

A space is *sober* if the correspondence $x \mapsto \mathfrak{N}_x$ is a bijection between points and completely prime filters. For sober spaces, we might just as well consider them as locales—we lose nothing by using the geometric theories to study sober topological spaces.

PROPOSITION 8.16 *Let X and Y be sober spaces. Then there is a bijection between maps $f : X \to Y$, and maps $g : [\mathrm{Th}_{\Omega X}] \to [\mathrm{Th}_{\Omega Y}]$ (homomorphisms $g^* : \Omega Y \to \Omega X$).*

Proof Apply Proposition 8.14 with $[\mathrm{Th}_{\Omega Y}]$ substituted for the locale Y. Sobriety assures us that the sober space Y is homeomorphic to $\mathrm{pt}([\mathrm{Th}_{\Omega Y}])$. QED

Any "point-free" approach, constructing the points out of the logic, is inevitably sober. Many well-behaved spaces, for instance all Hausdorff spaces, are sober, and any space can be "soberified" by replacing it by the space of completely prime filters.

The second mismatch between spaces and locales is that *not every locale comes from a space*. This arises out of an important logical fact, that *geometric logic is not complete.*

Stone's Theorem showed how each Boolean algebra is isomorphic to a sub-Boolean-algebra of a powerset. In logical terms, there are always enough standard models to distinguish between inequivalent sentences. This is a consequence of completeness.

By contrast, locales do not always have enough global points to discriminate between the opens. For each locale X, the extent homomorphism $\mathrm{ext} : \Omega X \to \mathcal{P}\,\mathrm{pt}(X)$ defines a topology on $\mathrm{pt}(X)$. The locale X is *spatial* if ext is 1-1, but not all locales are spatial.

For example, let $\mathbb{R}$ be the real line with its usual topology. Let T be $\mathrm{Th}_{\Omega\mathbb{R}}$ extended by extra axioms $\neg\neg U \to U$ for every $U \in \Omega\mathbb{R}$. ($\neg\neg$ is the Heyting

double negation in $\Omega\mathbb{R}$. Concretely, $\neg\neg U$ is the interior of the closure of U.) $\mathrm{pt}[T]$ is a subspace of $\mathbb{R}$, comprising those reals x such that for every U, if $x \in \neg\neg U$ then $x \in U$. There are no such x, for consider $U = (-\infty, x) \cup (x, \infty)$, which has $\neg\neg U = \mathbb{R}$. But $\neg\neg$ is a *nucleus* (see e.g. Johnstone, 1982), and an immediate consequence of the general theory is that the opens of $[T]$ are equivalent to the regular opens of $\mathbb{R}$, i.e. those U for which $\neg\neg U = U$. Hence $[T]$ is a non-trivial locale with no global points, hence non-spatial.

Logically, spatiality is the same as completeness, but there is a difference of emphasis. Completeness refers to the ability of the logical reasoning (from rules and axioms) to generate all the equivalences that are valid for the models: if not, then it is the logic that is considered incomplete. Spatiality refers to the existence of enough models to discriminate between logically inequivalent formulae: if not, then the class of models is incomplete.

In classical mathematics, most important locales are spatial; but this can rely on the axiom of choice to find sufficient points. In constructive mathematics many important locales (such as the real line) behave better in non-spatial form, and if we spatialize by topologizing the global points, then important theorems (such as the Heine-Borel Theorem) become false. This has led to a common misconception that constructive topology is deficient in theorems. This is actually not true, and the purpose of this chapter is to show how topology and constructive reasoning are intimately related. However, an important step is to forego any dependence on spatiality, on relying on a space being carried by an untopologized *set* of points.

Happily, constructive reasoning itself contains the key to dealing with non-spatiality. Spatiality is an issue when we try to deal with a locale in terms of its global points, of which there might not be enough. But there are enough generalized points. For example, a map of locales is defined by its action on the generic point. The generalized points live in non-Boolean lattices (or, as we shall shortly see, in the non-classical mathematics of sheaves), and it is convenient to deal with them using constructive mathematics as a tool.

2.5 Example: the localic reals

As an adaptation of the localic reals in Johnstone, 1982, we present a propositional geometric theory $T_{\mathbb{R}}$ with propositional symbols $P_{q,r}$ ($q, r \in \mathbb{Q}$, the rationals) and axioms:

$$P_{q,r} \wedge P_{q',r'} \leftrightarrow \bigvee\{P_{s,t} \mid \max(q,q') < s < t < \min(r,r')\}$$
$$\top \rightarrow \bigvee\{P_{q-\varepsilon,q+\varepsilon} \mid q \in \mathbb{Q}\} \text{ if } 0 < \varepsilon \in \mathbb{Q}$$

PROPOSITION 8.17 *In the theory $T_{\mathbb{R}}$, we can derive the following.*

$$P_{q,r} \vdash\dashv \bigvee\{P_{s,t} \mid q < s < t < r\}$$
$$P_{q,r} \wedge P_{q',r'} \vdash\dashv P_{\max(q,q'),\min(r,r')}$$
$$P_{q,r} \vdash \bot \qquad \textit{if } r \leq q$$
$$P_{q',r'} \vdash P_{q,r} \qquad \textit{if } q \leq q' \textit{ and } r' \leq r$$
$$P_{q,t} \vdash P_{q,r} \vee P_{s,t} \qquad \textit{if } q < s < r < t$$

Proof These are all straightforward except the last. If $q < s < r < t$ then let $\varepsilon = (r - s)/2$. We have

$$\begin{aligned} P_{q,t} &\vdash \bigvee\{P_{u-\varepsilon,u+\varepsilon} \mid u \in \mathbb{Q}\} \wedge P_{q,t} \\ &\vdash\dashv \bigvee\{P_{\max(u-\varepsilon,q),\min(u+\varepsilon,t)} \mid u \in \mathbb{Q}\}. \end{aligned}$$

Now for any $u \in \mathbb{Q}$, we cannot have both $u + \varepsilon > r$ and $u - \varepsilon < s$, for then $r - \varepsilon < u < s + \varepsilon$, so $r - s < 2\varepsilon$, contradiction. Hence either $u + \varepsilon \leq r$, in which case $P_{\max(u-\varepsilon,q),\min(u+\varepsilon,t)} \vdash P_{q,r}$, or $u - \varepsilon \geq s$, in which case $P_{\max(u-\varepsilon,q),\min(u+\varepsilon,t)} \vdash P_{s,t}$. QED

In Sec. 4.7 we shall see how the models of this are equivalent to Dedekind sections of the rationals.

We have a model of the theory in $\Omega\mathbb{R}$, interpreting $P_{q,r}$ as the open interval $(q, r) = \{x \in \mathbb{R} \mid q < x < r\}$, and hence a frame homomorphism $\alpha : \Omega[T_{\mathbb{R}}] \to \Omega\mathbb{R}$. Clearly it is onto, since the open intervals (q, r) form a base of opens. It is also 1-1, for suppose ϕ and ψ are elements of $\Omega[T_{\mathbb{R}}]$ (geometric formulae) such that $\alpha(\phi) \subseteq \alpha(\psi)$. We show that $\phi \vdash \psi$. Any finite meet of symbols $P_{q,r}$ is a join of such symbols, and it follows that any formula ϕ is equivalent to a join of such symbols. Hence it suffices to show that if $\alpha(P_{q,r}) \subseteq \alpha(\psi)$ then $P_{q,r} \vdash \psi$. From Proposition 8.17, it suffices to show that if $\alpha(P_{q,r}) \subseteq \alpha(\psi)$ and $q < q' < r' < r$ then $P_{q',r'} \vdash \psi$. Let $S = \{s \in \mathbb{Q} \mid q' \leq s \leq r'$ and $P_{q',s} \vdash \psi\}$. S is non-empty (because $q' \in S$) and bounded above (by r'), and so it has a supremum, a real number x. Since $q' \leq x \leq r'$ we have $x \in \alpha(P_{q,r})$ and so $x \in \alpha(\psi)$. Since ψ too is a join of symbols $P_{t,u}$, we can find one such that $P_{t,u} \vdash \psi$ and $x \in \alpha(P_{t,u})$, i.e. $t < x < u$. If $t \leq q'$, then $P_{q',u} \vdash \psi$. On the other hand, suppose $q' < t$. Choose a rational t' with $t < t' < x$. Then by definition of x we have $P_{q',t'} \vdash \psi$. Again, but this time using Proposition 8.17, we get $P_{q',u} \vdash \psi$. It follows that $\min(r', u) \in S$, so $x \leq \min(r', u) \leq x$. Since $x < u$, it follows that $x = r' < u$, so $P_{q',r'} \vdash \psi$.

It follows that the locale $[T_{\mathbb{R}}]$ is spatial, with $\Omega[T_{\mathbb{R}}] \cong \Omega\mathbb{R}$. However, the proof just given is classical, in particular in its assumption that a non-empty set of rationals, bounded above, has a real-valued supremum. There are

(Fourman and Hyland, 1979) non-classical examples where $[T_{\mathbb{R}}]$ is not spatial. This might be seen as a defect of the locale $[T_{\mathbb{R}}]$, but in fact this non-spatial locale has better constructive behaviour than the space $\mathbb{R}$. For example, the Heine-Borel Theorem holds for the locale but not, in general, for the space (Fourman and Grayson, 1982).

3. Predicate geometric logic

In a sense, the propositional geometric logic is all that is needed for treating "topologies as Lindenbaum algebras". However, there are good reasons for extending these ideas to the case of predicate logic.

The first is Grothendieck's discovery that there are certain situations that involve topology and continuity, but where topological spaces are inadequate for expressing them. He invented toposes to cover these situations, and said, "toposes are generalized topological spaces". The generalization is essentially that from propositional to predicate geometric logic, and the toposes are a categorical version of Lindenbaum algebra appropriate to this predicate case. Thus it would be more accurate to say that toposes are generalized locales. Again we can view the topos as a space of models, but in general there are not enough opens to define the generalized topological structure, and sheaves must be used instead. Another point of generalization is that the classes $\mathrm{Map}(X, Y)$ are no longer posets but have category structure—the specialization order is replaced by specialization morphisms. (Specialization morphisms between points correspond to homomorphisms between models.)

A second reason for studying the predicate logic is that quite often it is natural to replace a propositional geometric theory by an equivalent predicate theory. The reason this is possible in any but the most trivial cases is a remarkable consequence of having infinitary disjunctions. These allow sorts to be characterized uniquely up to isomorphism as, for example, the natural numbers. This means that there is an intrinsic type theory in predicate geometric logic, and one can work not so much in geometric *logic* as in a geometric *mathematics*, which turns out to have an intrinsic continuity.

To fix our logical terminology, we say that a *many-sorted, first-order signature* has a set of sorts, a set of predicate symbols, and a set of function symbols. Each predicate or function symbol has an *arity* stipulating the number and sorts of its arguments, and (for a function) the sort of its result. A predicate symbol with no arguments is *propositional*, while a function with no arguments is a *constant*. We shall express the arities of predicates and functions thus:

$P \subseteq A_1, ..., A_n$	(for a predicate)
$P \subseteq 1$	(for a proposition)
$f : A_1, ..., A_n \to B$	(for a function)
$c : B$	(for a constant)

We shall also freely use vector notation, writing e.g. $\vec{A}$ instead of $A_1, \ldots, A_n$. In many situations, as here, this is to be understood as representing a product.

DEFINITION 8.18 *Let Σ be a many-sorted, first-order signature. If $\vec{x}$ is a (finite) vector of distinct variables, each with a given sort, then a* geometric formula *over Σ in* context *$\vec{x}$ is a formula built up using term formation from the variables $\vec{x}$ and the function symbols of Σ, and formula formation from the terms and the predicate symbols from Σ using $=$, $\wedge$, $\top$, $\bigvee$ (possibly infinitary) and $\exists$. Note that, even with infinitary disjunctions, a formula is allowed only finitely many free variables, since they all have to be taken from the finite context $\vec{x}$. Not all the variables in the context have to be used in the formula.*

A geometric theory *over Σ is a set of axioms of the form*

$$(\forall \vec{x})\ (\phi \to \psi)$$

where ϕ and ψ are geometric formulae over Σ in context $\vec{x}$. (We shall also commonly write such axioms in sequent form $\phi \vdash_{\vec{x}} \psi$.)

A geometric theory is coherent *if all disjunctions used in it are finitary. (Note that Mac Lane and Moerdijk, 1992, X.3 uses the word "geometric" to mean "coherent".)*

When we need to make explicit reference to the context of a term or formula, we shall use notation such as $(\vec{x}.t)$ or $(\vec{x}.\phi)$.

DEFINITION 8.19 *Let T_1 and T_2 be two geometric theories. A* theory morphism *F from T_1 to T_2 comprises the following data.*

1. *To each sort A of T_1, there is assigned a sort $F(A)$ of T_2. After this, each arity α for T_1 can be translated to an arity $F(\alpha)$ for T_2.*

2. *To each function symbol f of T_1, with arity α, there is assigned a function symbol $F(f)$ of T_2, with arity $F(\alpha)$. After this, each term in context $(\vec{x}.t)$ for T_1 can be translated to a term in context $(\vec{x}.F(t))$ for T_2.*

3. *Similarly, to each predicate symbol P of T_1, with arity α, there is assigned a predicate symbol $F(P)$ of T_2, with arity $F(\alpha)$. After this, each formula in context $(\vec{x}.\phi)$ for T_1 can be translated to a formula in context $(\vec{x}.F(\phi))$ for T_2.*

4. *To each axiom $(\forall \vec{x})\ (\phi \to \psi)$ of T_1, there is an axiom $(\forall \vec{x})\ (F(\phi) \to F(\psi))$ of T_2.*

Note—theory morphisms are presentation-dependent, and do not provide the general notion of map.

3.1 Logical rules

For the logical rules of predicate geometric logic we again follow the account in Johnstone, 2002b, D1.3.1. They are expressed using sequents of the form $\phi \vdash_{\vec{x}} \psi$ where ϕ and ψ are formulae in context $\vec{x}$.

Labelling the turnstile with the context allows us to give a clean treatment of empty carriers, following Mostowski (see Lambek and Scott, 1986). This is exemplified by the following non-geometric deduction:

$$\begin{aligned}&(\forall x)\ \phi(x)\\&\phi(a)\\&(\exists x)\ \phi(x).\end{aligned}$$

This purports to prove a sequent $(\forall x)\ \phi(x) \vdash (\exists x)\ \phi(x)$, but that is invalid with an empty carrier. The true conclusion is that we can make the inference *in the context* a, which we write $(\forall x)\ \phi(x) \vdash_a (\exists x)\ \phi(x)$. This is valid provided a is interpreted. But from that we can not infer $(\forall x)\ \phi(x) \vdash (\exists x)\ \phi(x)$. (Some such device is necessary in constructive logic, where excluding empty carriers would be a serious problem. But it ought also to be better known in classical logic. See also Example 8.27.)

The rules of predicate geometric logic are those of the propositional logic (with the context labels added) together with the following: *substitution* is

$$\frac{\phi \vdash_{\vec{x}} \psi}{\phi(\vec{s}/\vec{x}) \vdash_{\vec{y}} \psi(\vec{s}/\vec{x})}$$

where $\vec{s}$ is a vector of terms in context $\vec{y}$, with sorts matching those of $\vec{x}$; the *equality* rules

$$\top \vdash_x x = x, \qquad (\vec{x} = \vec{y}) \wedge \phi \vdash_{\vec{z}} \phi(\vec{y}/\vec{x})$$

($\vec{z}$ has to include all the variables in $\vec{x}$ and $\vec{y}$, as well as those free in ϕ); the *existential* rules

$$\frac{\phi \vdash_{\vec{x},y} \psi}{(\exists y)\phi \vdash_{\vec{x}} \psi}, \qquad \frac{(\exists y)\phi \vdash_{\vec{x}} \psi}{\phi \vdash_{\vec{x},y} \psi};$$

and the *Frobenius* rule

$$\phi \wedge (\exists y)\psi \vdash_{\vec{x}} (\exists y)(\phi \wedge \psi).$$

Note that the substitution rule justifies *context weakening*

$$\frac{\phi \vdash_{\vec{x}} \psi}{\phi \vdash_{\vec{x},y} \psi}.$$

In other words, a deduction in one context will still be valid if we add extra variables, though not if we remove unused variables (which is what was done in the example of $(\forall x)\ \phi(x) \vdash (\exists x)\ \phi(x)$).

Syntax	Interpretation
sort A	*carrier* set $\{M\|A\}$
sort tuple $\vec{A} = (A_1, \ldots A_n)$	$\{M\|\vec{A}\} = \prod_{i=1}^{n}\{M\|A_i\}$
predicate $P \subseteq \vec{A}$	subset $\{M\|P\} \subseteq \{M\|\vec{A}\}$
proposition $P \subseteq 1$	subset $\{M\|P\} \subseteq 1$
function $f : \vec{A} \to B$	function $\{M\|f\} : \{M\|\vec{A}\} \to \{M\|B\}$
constant $c : B$	element $\{M\|c\} \in \{M\|B\}$
formula in context $(\vec{x}.\phi)$	subset $\{M\|\vec{x}.\phi\} \subseteq \{M\|\sigma(\vec{x})\}$
term in context $(\vec{x}.t)$	function $\{M\|\vec{x}.t\} : \{M\|\sigma(\vec{x})\} \to \{M\|\sigma(t)\}$

Figure 8.1. Interpretations of syntactic elements.

3.2 Models

The notion of standard model (in sets) is as expected, except that we *allow empty carriers*. The logical rules, with the context attached to the turnstile, are designed to be sound for empty carriers. We shall also introduce a novel notation that is useful when dealing with more than one interpretation at the same time.

The interpretation of different syntactic elements is defined in Table 8.1. The notation $\sigma(t)$ denotes the sort of a term t, and similarly for a tuple of terms. Once the signature ingredients are interpreted (arbitrarily), the interpretation of terms and formulae in context follows structurally in an evident way, so that $\{M|\vec{x}.\phi\}$ is the set of value tuples (in $\{M|\vec{A}\}$) for which ϕ holds, and $\{M|\vec{x}.t\}$ yields a result in $\{M|\sigma(t)\}$ for any value tuple substituted for $\vec{x}$.

If T is a geometric theory over Σ, then we say that an interpretation M of Σ is a *model* of T if, for every axiom $\phi \vdash_{\vec{x}} \psi$ in T, we have $\{M|\vec{x}.\phi\} \subseteq \{M|\vec{x}.\psi\}$.

We also define a notion of homomorphism.

DEFINITION 8.20 *Let Σ be a signature and let M, N be two interpretations of Σ. Then a* homomorphism *$h : M \to N$ comprises a function $\{h|A\} : \{M|A\} \to \{N|A\}$ for each sort A, subject to the following conditions. We shall write $\{h|\vec{A}\}$ for the product function $\prod_i\{h|A_i\} : \{M|\vec{A}\} \to \{N|\vec{A}\}$. For each predicate $P \subseteq \vec{A}$ in Σ and for each function $f : \vec{A} \to B$ we require*

$$\{M|P\} \subseteq \{h|\vec{A}\}^{-1}(\{N|P\})$$
$$\{h|B\} \circ \{M|f\} = \{N|f\} \circ \{h|\vec{A}\}.$$

Informally, we may say for any suitable value tuples $\vec{a}$ in M, that if $P(\vec{a})$ holds in M, then $P(h(\vec{a}))$ holds in N; and that $f(h(\vec{a})) = h(f(\vec{a}))$.

The two conditions, for predicates and functions, are not independent. If the function f is instead described by its graph, then the predicate condition for the graph is equivalent to the function condition for f.

Obviously homomorphisms can be composed, and there are identity homomorphisms, and so for any theory T we have a category $\mathrm{Mod}(T)$ of models of T.

PROPOSITION 8.21 *Let Σ be a signature, let M, N be two interpretations of Σ, and let $h : M \to N$ be a homomorphism.*

1 *Let $(\vec{x}.t)$ be a term in context. Then*

$$\{N|\vec{x}.t\} \circ \{h|\sigma(\vec{x})\} = \{h|\sigma(t)\} \circ \{M|\vec{x}.t\}.$$

2 *If $(\vec{x}.\phi)$ is any geometric formula in context, then $\{h|\sigma(\vec{x})\}$ restricts to a function*

$$\{h|\vec{x}.\phi\} : \{M|\vec{x}.\phi\} \to \{N|\vec{x}.\phi\}.$$

Proof Induction on the formation of terms and formulae. QED

The result relies fundamentally on the positivity of the logic. For a logic with negation, the homomorphism condition that we gave for predicates is not liftable through negation. For this reason in classical logic one may see a different notion of homomorphism.

REMARK 8.22 *Categorically, Definition 8.20 amounts to saying that for sorts B, sort tuples $\vec{A}$ and predicates $P \subseteq \vec{A}$, we have functors $|B\}$, $|\vec{A}\}$ and $|P\} : \mathrm{Mod}(T) \to \mathbf{Set}$ with natural transformations $|P\} \to |\vec{A}\}$; and for functions $f : \vec{A} \to B$ there are natural transformations $|f\} : |\vec{A}\} \to |B\}$. The Proposition says this extends to formulae and terms, with natural transformations $|\vec{x}.\phi\} \to |\sigma(\vec{x})\}$ and $|\vec{x}.t\} : |\sigma(\vec{x})\} \to |\sigma(t)\}$.*

If $F : T_1 \to T_2$ is a theory morphism, then for every model M of T_2 there is a corresponding model F^*M of T_1, defined by

$$\{F^*M|-\} = \{M|F(-)\}.$$

This is called the *F-reduct* of M. F^* gives a functor from $\mathrm{Mod}(T_2)$ to $\mathrm{Mod}(T_1)$ (note the reversed direction!).

3.3 Cartesian theories

These provide some important examples of geometric theories. They also provide the setting for some key constructions of universal algebra (including initial and free algebras) that turn out to be "geometric" in nature (Sec. 3.4).

The best known and simplest amongst the Cartesian theories are the finitary algebraic theories, where "finitary" refers to the requirement that all operators should have finite arity. Note that, unlike Johnstone, 2002b, Definition D1.1.7(a), we allow them to be many-sorted.

DEFINITION 8.23 *A* finitary algebraic theory *is a geometric theory presented with no predicate symbols, and with axioms all of the form*

$$(\forall \vec{x})\ (\top \to s = t)$$

where s and t are two terms in context $\vec{x}$. (In other words, the axioms are equational laws.*)*

Cartesian theories generalize these, essentially by allowing operators to be partial. In that generality they are slightly difficult to formalize and have appeared in various guises. The definition here (following Johnstone, 2002b, D1.3.4) is due to Coste. Equivalent are the essentially algebraic theories of Freyd, the left exact theories and sketches (see Barr and Wells, 1984) and the quasi-equational theories of Palmgren and Vickers, 2007.

As the references make clear, Cartesian theories are intimately associated with *Cartesian categories*, i.e. categories with all finite limits. A functor between them that preserves all finite limits is a *Cartesian functor*.

Most of the examples of Cartesian theories in this chapter are in fact finitary algebraic. Readers may safely omit the following definition if they wish, and consider Theorems 8.25 and 8.26 in the finitary algebraic case.

DEFINITION 8.24 (JOHNSTONE, 2002B, D1.3.4) *Let Σ be a many-sorted, first-order signature, and let T be a coherent theory over it.*

The formulae in context that are Cartesian *relative to T are as follows: atomic formulae, $\top$, equations and conjunctions of Cartesian formulae; and $(\vec{x}.(\exists y)\phi)$ provided $(\vec{x}, y.\phi)$ is Cartesian, and the following sequent is derivable from T:*

$$\phi \wedge \phi(y'/y) \vdash_{\vec{x}yy'} y = y'$$

The theory T is Cartesian if there is a well-founded partial order on its axioms, such that for every axiom $(\forall \vec{x})(\phi \to \psi)$, the formulae in context $(\vec{x}.\phi)$ and $(\vec{x}.\psi)$ are Cartesian relative to the previous axioms in T.

The essential point of this definition is that existential quantification can be used only when it is provably unique. This allows a mechanism for dealing with partial operations, by replacing them by their graphs. For example, the theory of categories is Cartesian but not algebraic. Composition is partial, with $f \circ g$ defined iff the domain of f is equal to the codomain of g.

As is well-known, every algebraic theory T has an initial model—that is to say, the category $\mathrm{Mod}(T)$ has an initial object. A consequence of this is that reduct functors have left adjoints. This generalizes to Cartesian theories.

THEOREM 8.25 (INITIAL MODEL THEOREM) *Let T be a Cartesian theory. Then T has an initial model, in other words a model M_0 such that for every other model M there is a unique homomorphism $M_0 \to M$.*

Proof In the case where T is an algebraic theory, the initial model is got by taking all terms, and then factoring out a congruence generated from the equational laws. (The construction of Lindenbaum algebras, as propositions modulo provable equality, is a particular instance of this.) Palmgren and Vickers, 2007 show how a similar proof can cover Cartesian theories, by working in a logic of partial terms. The construction first takes all partial terms, and then factors out a partial congruence (not necessarily reflexive) of provable equality, in which self-equality of a term is equivalent to its being defined.

More traditional proofs rely on first forming a *syntactic category*, a Cartesian category $\mathcal{C}_T$ such that models of T are equivalent to Cartesian functors from $\mathcal{C}_T$ to $\mathbf{Set}$ (Johnstone, 2002b, Theorem D1.4.7), and then appealing to Kennison's Theorem (Barr and Wells, 1984, Theorem 4.2.1). QED

THEOREM 8.26 (FREE MODEL THEOREM) *Let T_1 and T_2 be Cartesian theories, and let $F : T_1 \to T_2$ be a theory morphism. Then the reduct functor $F^* : \mathrm{Mod}(T_2) \to \mathrm{Mod}(T_1)$ has a left adjoint* Free_F.

Proof Just to sketch the proof, let M be a model of T_1. We can define a new Cartesian theory T whose models are pairs (N, f), where N is a model of T_2 and $f : M \to F^*(N)$ is a homomorphism. T is got by augmenting T_2 with constant symbols for the elements of M, and equations to say that the interpretation of those constants respects the structure of T_1. Then an initial model of T can be taken for $\mathrm{Free}_F(M)$. QED

The best known examples are where T_1 and T_2 are both single-sorted algebraic, with T_1 the theory with (one sort and) no operators or laws. Its models are sets. There is a unique theory morphism $F : T_1 \to T_2$, and F^* picks out the carrier but forgets all the algebraic structure. Then Free_F constructs the free T_2-model on a set.

Note that this theorem, and the initial model theorem on which it depends, in general rely critically on the fact that we allow empty carriers.

EXAMPLE 8.27 *Consider the algebraic theory with two sorts A and B, two constants s and t of sort B, a unary operator $f : A \to B$ and axioms*

$$(\forall x : A)\ (\top \to s = f(x))$$
$$(\forall x : A)\ (\top \to t = f(x)).$$

Its initial model M_0 has $\{M_0|A\} = \emptyset$, $\{M_0|B\} = \{s, t\}$.

Examples like this are sometimes used in equational reasoning to suggest that "equality is not transitive if empty carriers are allowed". This is because the equational laws can be presented as $s = f(x)$ and $t = f(x)$, but in M_0 we cannot deduce $s = t$. In our treatment we see that the equalities are in context, and in effect, "equality in context x" *is* transitive. We can deduce $\top \vdash_x s = t$, which implies $s = t$ *provided* we can interpret the variable x in A.

3.4 Geometric types

We have presented geometric theories using simple sorts that are declared in the signature and thereafter cannot be manipulated in any way. However, it would be an obvious convenience if we could perform mathematical constructions on those sorts to derive new ones. Following Johnstone, 2002b, D4.1 we shall use the word *type* for these generalized sorts, and reserve the word *sort* for what was declared in the signature.

An important feature of geometric logic is that its infinitary disjunctions allow us to characterize some type constructors uniquely up to isomorphism by using geometric structure and axioms.

A fundamental example is the natural numbers. Consider the geometric theory with a single sort N, constant 0, unary operator s and axioms

$$\begin{aligned}(\forall x : N)\ (s(x) = 0 \longrightarrow \bot)\\ (\forall x, y : N)\ (s(x) = s(y) \longrightarrow x = y)\\ (\forall x : N)\ \bigvee_{n\in\mathbb{N}} x = s^n(0).\end{aligned}$$

Here, $s^n(0)$ stands for the term $s(...(s(0))...)$ with n occurrences of s. The notation s^n is not a formal part of the logic, but a metasyntax used to describe the set of formulae over which the disjunction is taken. In any model, N can and must be interpreted as a set that is isomorphic to the natural numbers, by a unique isomorphism under which the constant 0 corresponds to the natural number 0, and the function s corresponds to the successor operation $n \longmapsto n+1$. Hence, modulo isomorphism, this theory is just a variant of the trivial theory with empty signature and no axioms.

In fact, just within the logic we can prove that N has a universal property that characterizes the natural numbers: it is an initial model for the single-sorted algebraic theory of *induction algebras*, which has constant ε, unary operator t and no axioms.

THEOREM 8.28 *Let a geometric theory have $N, 0, s$ as axiomatized above. Then, in any model, N is interpreted as an initial induction algebra.*

Proof Let (A, ε, t) be an induction algebra A. Of course N is itself an induction algebra under 0 and s; we must show there is a unique induction algebra homomorphism from N to A. For uniqueness, suppose $f : N \to A$ is a homomorphism. We show that (in sequent form)

$$y = f(x) \vdash\dashv_{xy} \bigvee_{n\in\mathbb{N}} (x = s^n(0) \wedge y = t^n(\varepsilon)).$$

For $\dashv$, we have $\top \vdash f(s^n(0)) = t^n(\varepsilon)$ by induction on n. For $\vdash$, combine this with

$$\top \vdash_x \bigvee_{n\in\mathbb{N}} (x = s^n(0)).$$

For existence of f, we show that the formula in context

$$(xy.\,\Gamma) \equiv \bigvee_{n\in\mathbb{N}} (x = s^n(0) \wedge y = t^n(\varepsilon))$$

is the graph of an induction algebra homomorphism, that is to say

$$\begin{aligned} \Gamma \wedge \Gamma(y'/y) &\vdash_{xyy'} y = y' \\ \top &\vdash_x (\exists y)\Gamma \\ \top &\vdash \Gamma(0, \varepsilon/x, y) \\ \Gamma &\vdash_{xy} \Gamma(s(x), t(y)/x, y). \end{aligned}$$

The first two of these state that the relation Γ is single-valued and total, and hence the graph of a function, and the remaining two state that the function preserves the induction algebra operations.

These are all easy except perhaps for the first. For that, we want

$$x = s^n(0) \wedge y = t^n(\varepsilon) \wedge x = s^{n'}(0) \wedge y' = t^{n'}(\varepsilon) \vdash_{xyy'} y = y'.$$

If $n = n'$ that is obvious, while for $n \neq n'$ we can prove $x = s^n(0) \wedge x = s^{n'}(0) \vdash \bot$. QED

Let us stress the fact that the proof was within the formality of geometric logic. Thus it will be valid not only for the standard semantics in sets, but also for other semantics such as (as we shall see) in sheaves.

As is well known, there can be no such characterization of the natural numbers in finitary logic. Our ability to do it in geometric logic derives from the power of the infinitary disjunctions, and this extends to other "geometric" type constructs.

One might respond to this power by regarding explicit geometric type constructions as unnecessary, since they can be reduced to first order logic. By contrast we shall instead feel justified in using an explicit geometric type theory, since it does not transcend the scope of geometric logic.

In the present state of our knowledge we do not have a formal type theory along these lines. Instead, we shall use types informally, using the geometric type constructs wherever sorts can occur, and also introducing any functions and predicates that are associated with those types. In Remark 8.34 we shall see a semantic characterization of which type constructs are geometric.

By *geometric type theory* we shall understand a geometric theory in which geometric type constructs are used in this way. They have the same expressive

power as geometric theories. An interpretation must interpret those types and the associated predicates and functions in the intended way.

A *coherent type theory* is a geometric type theory in which all disjunctions are finite. These are stronger in expressive power than coherent theories, since the type constructs may implicitly use infinitary disjunctions for their justification.

The first example of geometric type constructor, and one of the simplest, is the Cartesian product. (In fact all finite categorical limits are geometric.) Other examples arise out of the following general principle. Suppose $F : T_1 \to T_2$ is a theory morphism between Cartesian theories. Then the free model construction Free_F is geometric. Note that this will usually construct not only types, but also functions. In the following example, the principle can be applied with both theories algebraic (possibly many sorted) – exercise! But in each case it is also possible to specify the type geometrically and give a geometric proof of its universal property.

EXAMPLE 8.29 *1 The* natural numbers $\mathbb{N}$.

2 The list type A^* *over a type* A. *In* $\mathbf{Set}$ *its elements are finite lists of elements of* A. A^* *is the free monoid (having associative binary operation with a 2-sided identity element) over* A*; see Johnstone, 2002a, A2.5.15 for another treatment.*

3 Coproducts *(in* $\mathbf{Set}$*, disjoint unions).*

4 Coequalizers *(or, more particularly, quotients of equivalence relations).*

One very important type constructor is the *finite powerset* $\mathcal{F}$. This is discussed extensively in Johnstone, 2002b, D5.4 under the notation of K. (Constructively, the particular notion of finiteness being used is *Kuratowski finiteness*. The Kuratowski finite subsets of a set S are the elements of the $\cup$-subsemilattice of $\mathcal{P}S$ generated by the singletons. S is Kuratowski finite if it is a Kuratowski finite subset of itself.) It is a geometric construction because $\mathcal{F}A$ is the free semilattice over A. (A *semilattice* is a monoid $(A, 0, \vee)$ in which $\vee$ is *commutative* ($x \vee y = y \vee x$) and *idempotent* ($x \vee x = x$).) Johnstone, 2002b, D5.4 gives an alternative description analogous to that of list objects; see also Vickers, 1999.

A particularly important feature of the finite power type is that it enables us to internalize universal quantification, *provided* it is finitely bounded. Suppose $(x : A, y : B.\ \phi)$ is a formula in context. Then so is $(S : \mathcal{F}A, y : B.\ (\forall x \in S)\ \phi)$. It is interpreted as follows. Consider $\{M|xy.\phi\}$ as a function from $\{M|A\}$ to $\mathcal{P}\{M|B\}$. The codomain of this is a semilattice under $\cap$, and so we get a semilattice homomorphism from $\mathcal{F}\{M|A\}$ to $\mathcal{P}\{M|B\}$. This transposes to a subset of $\{M|\mathcal{F}A \times B\}$, the interpretation of $(\forall x \in S)\ \phi$.

3.5 Dedekind sections

Each real number x is characterized by its *Dedekind section*, two sets of rationals:

$$L = \{q \in \mathbb{Q} \mid q < x\},$$
$$R = \{r \in \mathbb{Q} \mid x < r\}.$$

(Variants of this are possible, with $\leq$ instead of $<$. But they do not yield a geometric theory.) The idea then is to *define* the real number x to be the pair (L, R) of subsets of $\mathbb{Q}$, and (for $q, r \in \mathbb{Q}$) *define* $q < x$ if $q \in L$ and $x < r$ if $r \in R$.

The pairs $x = (L, R)$ that arise in this way are characterized by the following properties.

1 There is some rational q with $q < x$.

2 If $q < q' < x$ then $q < x$.

3 If $q < x$ then there is some rational q' with $q < q' < x$.

4 There is some rational r with $x < r$.

5 If $x < r' < r$ then $x < r$.

6 If $x < r$ then there is some rational r' with $x < r' < r$.

7 It is impossible to have $q < x < q$.

8 If $q < r$ then either $q < x$ or $x < r$.

((8) looks more obvious contrapositively: if $r \leq x \leq q$ then $r \leq q$. But we are axiomatizing $<$, not $\leq$.)

We can rewrite this, though at some cost in clarity, to a coherent type theory Ded with two predicates $L, R \subseteq \mathbb{Q}$ and axioms

$$\begin{aligned}
&(\exists q : \mathbb{Q})\ L(q)\\
&(\forall q, q' : \mathbb{Q})\ (q < q' \wedge L(q') \longrightarrow L(q))\\
&(\forall q : \mathbb{Q})\ (L(q) \longrightarrow (\exists q' : \mathbb{Q})\ (q < q' \wedge L(q')))\\
&(\exists r : \mathbb{Q})\ R(r)\\
&(\forall r, r' : \mathbb{Q})\ (r' < r \wedge R(r') \longrightarrow R(r))\\
&(\forall r : \mathbb{Q})\ (R(r) \longrightarrow (\exists r' : \mathbb{Q})\ (r' < r \wedge R(r')))\\
&(\forall q : \mathbb{Q})\ (L(q) \wedge R(q) \longrightarrow \bot)\\
&(\forall q, r : \mathbb{Q})\ (q < r \longrightarrow L(q) \vee R(r))
\end{aligned}$$

To see that this is indeed a coherent type theory, we must show that $\mathbb{Q}$ is a geometric type construction, and in addition that $<$ (on rationals) is geometric. The standard construction of $\mathbb{Q}$ is in stages. First, the natural numbers $\mathbb{N}$ have already been mentioned as an example of a geometric type. The arithmetic operations of addition and multiplication can then be defined in the following way. Addition is the unique operation $+ : \mathbb{N} \times \mathbb{N}$ that satisfies

$$(\forall n : \mathbb{N})\ 0 + n = n$$
$$(\forall m, n : \mathbb{N})\ s(m) + n = s(m + n).$$

Hence if the symbol $+$ is declared and those axioms are added to the theory, there is no change to the models – the operation $+$ is forced to be interpreted in the intended way. So its use as a standard mathematical symbol is shorthand for that declaration with axioms. Similarly, we can define all the relations $=, \neq, <, >, \leq$ and $\geq$ as subsets of $\mathbb{N} \times \mathbb{N}$. (In this case, we can even define them as operations $\mathbb{N} \times \mathbb{N} \to 2 = 1 + 1$. This is because the relations are decidable.) For instance, $<$ is the unique relation satisfying

$$(\forall n : \mathbb{N})\ 0 < s(n)$$
$$(\forall m : \mathbb{N})\ (m < 0 \to \bot)$$
$$(\forall m, n : \mathbb{N})\ (s(m) < s(n) \to m < n)$$
$$(\forall m, n : \mathbb{N})\ (m < n \to s(m) < s(n)).$$

Next, the integers $\mathbb{Z}$ are got as a quotient of $\mathbb{N} \times \mathbb{N}$ by an equivalence relation $\sim_1$, defined by $(m, n) \sim_1 (m', n')$ iff $m + n' = m' + n$. The pair (m, n) represents the integer $m - n$. Again, it is possible to define arithmetic and inequalities.

Finally, the rationals $\mathbb{Q}$ are got as a quotient of $\mathbb{Z} \times \{n \in \mathbb{N} \mid n \neq 0\}$ by an equivalence relation $\sim_2$, defined by $(p, q) \sim_2 (p', q')$ iff $pq' = p'q$. (The pair (p, q) represents the rational p/q.) The inequality $<$ is defined by $(p, q) < (p', q')$ if $pq' < p'q$.

So we see that $\mathbb{Q}$ and much of its accompanying structure are all geometric and can be used as needed in coherent type theories. This is emphatically *not* the case with the reals $\mathbb{R}$. As a set, $\mathbb{R}$ is described as a subset of $\mathcal{P}\mathbb{Q} \times \mathcal{P}\mathbb{Q}$, and the powerset constructor $\mathcal{P}$ is *not* amongst the geometric type constructors. (See Remark 8.34.) That is why we have to access the reals in a different way, by defining a theory whose models they are.

In Sec. 2.5 we saw the localic reals, given by a propositional geometric theory. In fact (Sec. 4.7), that geometric theory is equivalent to this one, despite the fact that one is propositional and the other is predicate with type constructs. In Sec. 4.6 we shall see how Ded retains a propositional character from the fact that it has no sorts declared. Its types are all constructed out of nothing.

4. Categorical logic

For propositional logic, the standard semantics interprets propositions as truth values. For a more general semantics, we interpreted propositions as elements of more general lattices—Boolean algebras for classical logic, frames for geometric logic. Then the Lindenbaum algebra was the lattice (of the appropriate kind) freely generated by a generic model.

For predicate logic, the standard semantics is in sets. Categorical logic generalizes this by interpreting the symbols in a category, of a kind appropriate to the logic, and then the analogue of the Lindenbaum algebra is a category. For geometric logic, the appropriate categories are Grothendieck toposes and the Lindenbaum algebra for a predicate geometric theory is the classifying topos. Our aim now is to explore how the technology of Lindenbaum algebras (as used for locales) extends to this setting.

We start with an introduction to categorical logic, following Johnstone, 2002b. A more elementary introduction can be found in Goldblatt, 1979.

4.1 Interpreting logic in a category

We assume that the reader has an elementary knowledge of category theory, including the basic definition, limits and colimits, and adjunctions.

We shall normally write composition of morphisms in applicative order, using "$\circ$". However, on occasion it will be convenient to use diagrammatic order instead, with ";". Thus the composition of morphisms

$$\xrightarrow{f}\xrightarrow{g}$$

will be written usually as $g \circ f$, but occasionally as $f; g$.

Suppose Σ is a many-sorted, first-order signature. The usual notion of interpretation of Σ in the category $\mathbf{Set}$ of sets, as given in Table 8.1, can be extended to *any* category $\mathcal{C}$ with finite products. Sets, functions, elements and subobjects become objects, morphisms, morphisms with domain 1 (a terminal object, i.e. nullary product) and subobjects in $\mathcal{C}$.

If A is an object in $\mathcal{C}$, we shall write $\mathrm{Sub}_{\mathcal{C}}(A)$ (or often just $\mathrm{Sub}(A)$) for the class of subobjects of A. We shall generally assume also that $\mathcal{C}$ is *well-powered*, i.e. that each $\mathrm{Sub}_{\mathcal{C}}(A)$ is a set. In our situations $\mathrm{Sub}_{\mathcal{C}}(A)$ is a meet semilattice, with greatest lower bounds given by pullbacks of subobjects. If $f : A \to B$ is a morphism, then pullback gives a meet semilattice homomorphism $f^* : \mathrm{Sub}_{\mathcal{C}}(B) \to \mathrm{Sub}_{\mathcal{C}}(A)$, called *inverse image*. (Exercise: in $\mathbf{Set}$ that is exactly what it is.)

As before, any interpretation of the ingredients of Σ in $\mathcal{C}$ can (given suitable categorical structure in $\mathcal{C}$) be extended recursively to terms and formulae in

context. We shall assume initially that $\mathcal{C}$ has at least all finite limits, in other words that it is *Cartesian*. First, we deal with terms.

Variables: $(x.x)$ is a term in context, interpreted by the identity morphism on $\{M|\sigma(x)\}$.

Substitution: Suppose $(\vec{x}.t)$ is a term in context, and $(\vec{w}.\vec{s})$ is a vector of terms in context that is type compatible with $\vec{x}$ (i.e. $\sigma(\vec{s}) = \sigma(\vec{x})$). (Note: we have the same context $\vec{w}$ for every component of $\vec{s}$.) Then $(\vec{w}.t(\vec{s}/\vec{x}))$ is a term in context. Its interpretation is given by

$$\langle\{M|\vec{w}.\vec{s}\}\rangle\,;\{M|\vec{x}.t\}:$$
$$\{M|\sigma(\vec{w})\} \to \{M|\sigma(\vec{x})\} \to \{M|\sigma(t)\}$$

where $\langle\{M|\vec{w}.\vec{s}\}\rangle$ denotes the product tupling $\langle\{M|\vec{w}.s_1\}, ..., \{M|\vec{w}.s_n\}\rangle$ of morphisms.

Substitution also covers *context weakening*. Suppose $(\vec{x}.t)$ is a term in context, and w is a variable not in $\vec{x}$. Then $(\vec{x}, w.\ \vec{x})$ is a vector of terms in context, and the substitution $(\vec{x}, w.\ t(\vec{x}/\vec{x}))$ gives a term in context $(\vec{x}, w.\ t)$. For its semantics, note that $\langle\{M|\vec{x}, w.\ \vec{x}\}\rangle$ is the product projection $\{M|\sigma(\vec{x})\} \times \{M|\sigma(w)\} \to \{M|\sigma(\vec{x})\}$.

Now we deal with formulae. Exercise! Check that these all make sense in $\mathbf{Set}$.

Substitution: Suppose $(\vec{x}.\phi)$ is a formula in context, and $(\vec{w}.\vec{t})$ is a vector of terms in context such that $\sigma(\vec{t}) = \sigma(\vec{x})$. Then $(\vec{w}.\phi(\vec{t}/\vec{x}))$ is a formula in context given by an inverse image (a pullback),

$$\{M|\vec{w}.\phi(\vec{t}/\vec{x})\} = \left\langle\{M|\vec{w}.\vec{t}\}\right\rangle^* (\{M|\vec{x}.\phi\}).$$

Again, this also covers *context weakening*. If $(\vec{x}.\phi)$ is a formula in context, and w is a variable not in $\vec{x}$, then $(\vec{x}, w.\ \phi)$ is also a formula in context given by substituting $\vec{x}$ for $\vec{x}$.

Equality: Let $(\vec{x}.t)$ and $(\vec{x}.t')$ be two terms in context, with $\sigma(t) = \sigma(t')$. Then $(\vec{x}.\ t = t')$ is a formula in context interpreted by an equalizer

$$\{M|\vec{x}.\ t = t'\} \hookrightarrow \{M|\sigma(\vec{x})\} \underset{\{M|\vec{x}.t'\}}{\overset{\{M|\vec{x}.t\}}{\rightrightarrows}} \{M|\sigma(t)\}.$$

Conjunction: Let $(\vec{x}.\phi)$ and $(\vec{x}.\psi)$ be two formulae in context. Then the conjunction rules imply that $(\vec{x}.\phi \wedge \psi)$ must be interpreted by the greatest lower bound of subobjects $\{M|\vec{x}.\phi\}$ and $\{M|\vec{x}.\psi\}$ in $\{M|\sigma(\vec{x})\}$. For $\top$, the nullary conjunction, we have $\{M|\vec{x}.\top\} = \{M|\sigma(\vec{x})\}$.

For the remaining geometric connectives, $\bigvee, \bot, \exists$, we need extra structure on the category $\mathcal{C}$.

For *existential quantification*, $\mathcal{C}$ must have *images*: the image of $f : A \to B$, if it exists, is the smallest subobject of B through which f factors. A

consequence of $\mathcal{C}$ having images (for all morphisms) is that there are "direct image" functions $\exists_f : \mathrm{Sub}(A) \to \mathrm{Sub}(B)$, left adjoint to f^*. (In **Set** the adjunction appears as $f(S) \subseteq T$ iff $S \subseteq f^{-1}(T)$.) As was first understood by Lawvere, this is exactly the content of the existential rules. Hence if $(\vec{x}, y.\ \phi)$ is a formula in context, so is $(\vec{x}.\ (\exists y)\ \phi)$ and it is interpreted by

$$\{M|\vec{x}.\ (\exists y)\ \phi\} = \exists_\pi(\{M|\vec{x}, y.\ \phi\})$$

where $\pi : \{M|\sigma(\vec{x})\} \times \{M|\sigma(y)\} \to \{M|\sigma(\vec{x})\}$ is the product projection.

For *disjunction,* the disjunction rules imply that if $(\vec{x}.\phi_i)$ are formulae in context, then $\{M|\vec{x}.\bigvee_i \phi_i\}$ has to be the least upper bound of the subobjects $\{M|\vec{x}.\phi_i\}$ in $\{M|\sigma(\vec{x})\}$. Hence our categorical structure must include those least upper bounds. *False* is just a nullary disjunction. $\{M|\vec{x}.\bot\}$ must be the least subobject of $\{M|\sigma(\vec{x})\}$.

Negation, implication, universal quantification: These connectives enter geometric logic only at the level of axioms, and we interpret them in a different way. Suppose we have formulae in context $(\vec{x}.\phi)$ and $(\vec{x}.\psi)$. Then an interpretation M *satisfies the axiom* $(\forall\vec{x})\ (\phi \to \psi)$, symbolically

$$M \vDash (\forall\vec{x})\ (\phi \to \psi),$$

if $\{M|\vec{x}.\phi\} \leq \{M|\vec{x}.\psi\}$. Negation can also be treated this way, by taking $\neg\phi$ as $\phi \to \bot$.

As usual, if T is a theory over signature Σ, then M is a model of T if it satisfies every axiom in T.

At this point it is very useful to consider the notion of *generalized element*, analogous to generalized points.

DEFINITION 8.30 *Let $\mathcal{C}$ be a category and A an object in it. A* generalized element *of A is any morphism whose codomain is A. The domain of the morphism is called the* stage of definition *of the generalized element.*

A generalized element at stage 1 *is a* global element.

Consider for instance the above assertion $M \vDash (\forall\vec{x})\ (\phi \to \psi)$. We might like this to mean that every element of $\{M|\sigma(\vec{x})\}$ that satisfies ϕ (i.e. it factors via $\{M|\vec{x}.\phi\}$) also satisfies ψ. But this is only a weak assertion if there is a shortage of morphisms from 1 to $\{M|\sigma(\vec{x})\}$.

Our interpretation of the assertion $M \vDash (\forall\vec{x})\ (\phi \to \psi)$ can now be explained naturally in terms of generalized elements, for it says that every *generalized* element of ϕ is also in ψ. To see this one way round, consider the inclusion $\{M|\vec{x}.\phi\} \to \{M|\sigma(\vec{x})\}$. This is a generalized element (with stage of definition $\{M|\vec{x}.\phi\}$) that satisfies ϕ. In fact it is the *generic* element of ϕ in M. If it is also to satisfy ψ, then we get $\{M|\vec{x}.\phi\} \leq \{M|\vec{x}.\psi\}$. Conversely, every generalized element of ϕ factors through the generic element, so if this is in ψ so too is every generalized element of ϕ.

Normally, when we intend our language to be interpreted in categories like this then by "element" we shall mean generalized element.

Homomorphisms between interpretations are defined just as in Definition 8.20, modulo obvious changes—the carrier functions $\{f|A\}$ become morphisms, $\{h|\vec{A}\}^{-1}$ becomes $\{h|\vec{A}\}^*$, subset inclusion $\subseteq$ becomes subobject inclusion $\leq$, etc. Proposition 8.21 still holds, by induction on the formation of terms and formulae.

Again we have a category $\mathrm{Mod}_{\mathcal{C}}(T)$ of models of T in $\mathcal{C}$, and for any theory morphism $F : T_1 \to T_2$, we have an F-*reduct* functor $F^* : \mathrm{Mod}_{\mathcal{C}}(T_2) \to \mathrm{Mod}_{\mathcal{C}}(T_1)$.

4.2 Grothendieck toposes

To interpret geometric logic categorically, we shall use *Grothendieck toposes*. These are usually defined as "categories of sheaves over Grothendieck topologies", but that is in effect referring to a representation theorem, Giraud's Theorem (8.67). This says that a category is equivalent to such a category of sheaves iff it has certain structure and properties. Since the structure and properties can be related directly to the geometric logic, we shall use it as our definition. It can be described in various ways; our presentation here is the ∞-*pretopos with separating set of objects* of Johnstone, 2002b, Theorem C2.2.8 (vii) and Johnstone, 2002a, A1.4.

DEFINITION 8.31 *A category $\mathcal{E}$ is a* Grothendieck topos *if it has the following properties.*

1 *$\mathcal{E}$ has all finite limits.*

2 $\mathcal{E}$ has images *(p. 458) and image factorization is preserved by pullback.*

3 *$\mathcal{E}$ is* well-powered *(i.e. for every object A, the class $\mathrm{Sub}_{\mathcal{E}}(A)$ is a set).*

4 *For each object A, the poset $\mathrm{Sub}_{\mathcal{E}}(A)$ has arbitrary joins (least upper bounds), and they are preserved by pullbacks f^*.*

5 *Every set-indexed family of objects of $\mathcal{E}$ has a disjoint coproduct. (A coproduct $A = \sum_i A_i$ is* disjoint *if all the coproduct injections $A_i \to A$ are monic, and the meet of A_i and A_j in $\mathrm{Sub}_{\mathcal{E}}(A)$ is less than $\bigvee\{A_k \mid k = i \text{ and } k = j\}$. Hence if $i \neq j$ then A_i and A_j are disjoint subobjects of $\mathrm{Sub}_{\mathcal{E}}(A)$.)*

6 *Every equivalence relation $R \rightrightarrows A$ in $\mathcal{E}$ is a kernel pair. (Fuller details can be found in Mac Lane and Moerdijk, 1992, Appendix, Theorem 1 or Johnstone, 2002a, A1.3.6. If a pair of morphisms $a, b : R \to A$ is such that $\langle a, b\rangle : R \to A \times A$ is monic, then they can be thought of as a relation*

on A. Then the usual notions of reflexive, symmetric and transitive can be translated into categorical terms by the usual logical interpretation. Our condition then says there is a quotient *morphism $q : A \to B$ such that (a, b) is the kernel pair of q, i.e. they complete the pullback square of q pulled back against itself.)*

7 *$\mathcal{E}$ has a* separating *set S (not a proper class) of objects (i.e. if $f, g : A \to B$ are such that for every $u : C \to A$ with $C \in S$ we have $f \circ u = g \circ u$, then $f = g$).*

Condition (1) says that $\mathcal{E}$ is *Cartesian*, and enables us to interpret the logic of conjunction and equality, as well as substitution.

Adding condition (2) makes $\mathcal{E}$ *regular* (Johnstone, 2002a, A1.3.3) and enables us to interpret the logic of $\exists$; preservation under pullback gives the Frobenius rule. It also enables the technique (e.g. Theorem 8.28) of defining a morphism by its graph.

Adding conditions (3) and (4) makes $\mathcal{E}$ *geometric* (Johnstone, 2002a, A1.4.18) and enables us to interpret the logic of arbitrary disjunction; preservation under pullback gives the frame distributivity rule. This structure is already sufficient for interpreting all the first-order part of geometric logic.

Adding condition (5) makes $\mathcal{E}$ *∞-positive*, and then adding (6) makes it an *∞-pretopos* (Johnstone, 2002a, A1.4.19). These enable us to interpret the geometric type theory. In particular, the Initial and Free Model Theorems 8.25 and 8.26 still work for models of Cartesian theories in $\mathcal{E}$. The proofs can be checked to go through; more explicitly, the proof in Palmgren and Vickers, 2007 is valid in Heyting pretoposes, and that includes Grothendieck toposes.

Condition (7) is a "smallness" condition. It allows us to deduce—in Giraud's Theorem—that although $\mathcal{E}$ is large, it can still be generated from a small structure.

We wish to find the appropriate notion of Lindenbaum algebra, in the form of a Grothendieck topos, for a geometric theory. This will be called the *classifying topos* for the theory. The central result for an ordinary Lindenbaum algebra was Proposition 8.3. We shall replace "Boolean algebra" by "Grothendieck topos", but we also need to know the appropriate notion of "homomorphism of Grothendieck topos".

All the structure needed for geometric logic and type theory (including the image factorization and the joins of subobjects) can be constructed using finite limits and arbitrary colimits, and conversely those are geometric type constructs (characterizable uniquely up to isomorphism by geometric structure and axioms). We shall therefore be interested in functors between toposes that preserve colimits and finite limits. A functor preserves colimits if it has a right adjoint, and for Grothendieck toposes the converse can also be shown

(Johnstone, 2002b, Remark C2.2.10). A *geometric morphism* is an adjoint pair of functors between toposes for which the left adjoint (which preserves all colimits) preserves finite limits.

At this point we are going to introduce some non-standard notation arising out of the fundamental split personality of toposes—spatial (generalized spaces) or logical (generalized universes of sets). The real interest of geometric morphisms is that they are the topos analogue of continuous map: they are a notion from the spatial side. In fact, we shall often refer to them as maps. On the other hand the Grothendieck toposes as we know them up till now, the categories with structure that was used to interpret logic, are very much the generalized universes of sets. We shall introduce notation that distinguishes between the two sides in the same way as we distinguished between locales and frames. Thus although technically we are dealing with those categories, we shall use notation that allows them to pretend to be spaces.

If we declare a symbol (e.g. X) to denote a topos, we shall nonetheless reserve its use for the topos in its spatial aspect. When we want to refer to it in its logical aspect, in other words the actual category, we shall write $\mathcal{S}X$. We shall call the objects of $\mathcal{S}X$ the *sheaves* over the topos rather than the objects of the topos. The symbol $\mathcal{S}$ can be read as standing for "sheaves".

We therefore define:

DEFINITION 8.32 *Let X and Y be two toposes. A* geometric morphism *(or* map*) $f : X \to Y$ is a pair of functors*

$$f^* : \mathcal{S}Y \to \mathcal{S}X$$
$$f_* : \mathcal{S}X \to \mathcal{S}Y$$

such that f^* is left adjoint to f_* and f^* preserves finite limits.

f^* is called the inverse image part*, and f_* the* direct image part.

We write $\mathrm{Map}(X, Y)$ *for the class of geometric morphisms from X to Y.*

Note the directions! The structure preserving functor is f^*, and this goes in the *reverse* direction to the geometric morphism f. As we shall see later, f^* is analogous to the inverse image function for a continuous map between topological spaces. (However, the "direct image part" is not analogous to the direct image function.)

PROPOSITION 8.33 *Let $f : X \to Y$ be a geometric morphism. Then f^* preserves free model constructions for Cartesian theories.*

Proof (See Johnstone, 2002b, D5.3.7 for the case of free algebras over sets for a single-sorted theory.) Let $\alpha : T_1 \to T_2$ be a theory morphism between two Cartesian theories. Let A be a T_1-algebra in $\mathcal{S}Y$, and let $h : A \to \alpha^*(T_2\langle A\rangle)$

be a free T_2-algebra over it. (h is a T_1-homomorphism.) It is required to show that $f^*(h) : f^*(A) \to f^*(\alpha^*(T_2\langle A\rangle))$ is a free T_2-algebra over $f^*(A)$.

For models of Cartesian theories, any functor that preserves finite limits will transform models to models. This applies to both f^* and f_*. They also both preserve model reduction α^*. Moreover, the adjunction of f^* and f_* extends to models: there is a bijection between homomorphisms $f^*(A) \to B$ in $\mathcal{S}X$ and homomorphisms $A \to f_*(B)$ in $\mathcal{S}Y$. If B is a reduct $\alpha^*(B')$, then we see that T_1-homomorphisms $f^*(A) \to B$ are equivalent to T_2-homomorphisms $T_2\langle A\rangle \to f_*(B')$ and hence to T_2-homomorphisms $f^*(T_2\langle A\rangle) \to B'$. Hence $f^*(T_2\,\langle A\rangle)$ is the free T_2-model over $f^*(A)$, as required. QED

REMARK 8.34 *We can now state a general semantic characterization of "geometric type construct". They are those constructs that can be carried out in any Grothendieck topos, and are preserved by inverse image functors of geometric morphisms. (Remember that those inverse image functors are the analogues of homomorphisms between Lindenbaum algebras.) Those we have seen include* finite limits *(in set-theoretic terms: products, singletons (as terminal object), fibred products (pullbacks) and equalizers);* arbitrary colimits *(disjoint unions, quotients);* images*;* free model constructions *for theory morphisms between Cartesian theories (including the natural numbers, finite powersets and list objects); and* integers $\mathbb{Z}$ *and* rationals $\mathbb{Q}$*, and associated structure including arithmetic and inequalities.*

These are type constructs that can be permitted, informally, in a geometric or coherent type theory.

Our next result is the analogue of the specialization order on maps between locales (Definition 8.10).

THEOREM 8.35 *Let X and Y be toposes. Then:*

1. $\mathrm{Map}(X, Y)$ *is a category. The morphisms are called* specialization morphisms*, or* natural transformations*. If α is a specialization morphism from f to g, then we write $\alpha : f \Rightarrow g$.*

2. *Composition with maps on either side is functorial.*

3. *Composition with maps satisfies the "interchange law". Suppose $\alpha : f \Rightarrow g$ in $\mathrm{Map}(X, Y)$, and $\beta : h \Rightarrow k$ in $\mathrm{Map}(Y, Z)$. Then the following diagram commutes.*

$$\begin{array}{ccc} h \circ f & \overset{\beta\circ f}{\to} & k \circ f \\ h \circ \alpha \downarrow & & \downarrow k \circ \alpha \\ h \circ g & \underset{\beta\circ g}{\to} & k \circ g \end{array}$$

This allows us to define a horizontal composition *$\beta \circ \alpha : h \circ f \Rightarrow k \circ g$.*

Proof 1. Mac Lane and Moerdijk, 1992, Sec. VII.1. A morphism α from f to g, is defined as a natural transformation from f^* to g^*. These are equivalent to natural transformations from g_* to f_*. (Note the reversal of direction.)

2, 3. These are obvious and come from horizontal composition of natural transformations. (Mac Lane, 1971) QED

An important feature of maps is that we can take filtered colimits. These are a categorical generalization of the directed joins of locale maps (Proposition 8.11).

DEFINITION 8.36 *Let $\mathcal{C}$ be a category. Then $\mathcal{C}$ is* filtered *if it satisfies the following conditions.*

1 *$\mathcal{C}$ has an object.*

2 *If A and B are objects of $\mathcal{C}$, then there is an object C with morphisms $f : A \to C$ and $g : B \to C$.*

3 *If A and B are objects of $\mathcal{C}$, and $f, g : A \to B$, then there is an object C and morphism $h : B \to C$ such that $h \circ f = h \circ g$.*

To put this more concisely, $\mathcal{C}$ is filtered iff every finite diagram in $\mathcal{C}$ has a cocone. A poset is filtered iff it is directed.

Composition with maps preserves filtered colimits, and this preservation of filtered colimits is an important property of maps, analogous to *Scott continuity*.

THEOREM 8.37 *Let X and Y be Grothendieck toposes. Then:*

1 $\mathrm{Map}(X, Y)$ *has all filtered colimits.*

2 *Composition with maps on either side preserves filtered colimits.*

Proof 1. Suppose we have a filtered diagram of maps f_i. Then $\mathrm{colim}_i\, f_i$ is calculated by

$$(\operatorname*{colim}_i f_i)^*(B) = \operatorname*{colim}_i (f_i^*(B))$$

Regardless of filteredness, this will preserve colimits. The filteredness ensures that it preserves finite limits, because filtered colimits commute with finite limits.

2. For precomposition by $g : W \to X$, this follows from the fact that g^* preserves colimits. For postcomposition by $h : Y \to Z$, it is trivial. QED

4.3 Elementary toposes

Grothendieck toposes have the structure needed to interpret geometric logic. Surprisingly, they also turn out to have structure for interpreting full first-order logic and even higher-order logic, though that structure is not geometric—it is not preserved by inverse image functors. (In this it is like the Heyting arrow, which exists in frames but is not preserved by frame homomorphisms.) This led to a generalized notion of topos, the *elementary topos*, which embodies the finitary part of that fuller structure.

This structure allows $\neg$, $\rightarrow$ and $\forall$ as connectives for constructing formulae, and so allows the coherent axioms to be formulae. It is only with this step that the differences between classical and constructive logic become visible. Characteristically classical axioms such as

$$\begin{array}{ll} \neg\neg\phi \rightarrow \phi & \text{(double negation rule)} \\ \phi \vee \neg\phi & \text{(excluded middle)} \end{array}$$

cannot be stated in coherent form, since they require negation to be used as a connective.

A minimal definition uses the notion of *powerobject* $\mathcal{P}(A)$, an object whose elements at stage B are the subobjects of $B \times A$ (so the global elements are the subobjects of A, like a powerset).

DEFINITION 8.38 *An* elementary topos *is a Cartesian category with a power object* $\mathcal{P}(A)$ *for every object* A.

The standard texts (Mac Lane and Moerdijk, 1992, IV.1, Johnstone, 2002a) show how much more structure can be deduced from this. In particular, an elementary topos is Cartesian closed: if A and B are objects, then there is a further object B^A, the *exponential*, whose elements at stage C are in bijection with the morphisms from $C \times A$ to B. It also has finite colimits, and the subobject pullback functions f^* have both left adjoints $\exists_f$—as required in Sec. 4.1 to interpret $\exists$—and right adjoints $\forall_f$—which are needed for $\forall$.

The powerobject $\mathcal{P}(1)$ is known as the *subobject classifier*, Ω. Its elements at stage A are the subobjects of A. In particular, its global elements are the subobjects of 1 and can be thought of as truth values. In $\mathbf{Set}$, $\Omega \cong 2$ where 2 is defined as the coproduct $1 + 1$. But, logically, this implies excluded middle, and does not hold in general. 2 is the object of *decidable* truth values.

Note that we obtain type constructors in elementary toposes that are *non*-geometric—not preserved by inverse image functors. Important examples include Ω, powersets, function sets (exponentials) and the set of reals.

The existence of the subobject classifier has a big effect on the way the logic is interpreted. Subobjects of A are now equivalent to their characteristic morphisms $A \rightarrow \Omega$, and so logical formulae can be interpreted as *terms of type*

Ω. This is formalized in the *Mitchell-Bénabou language* (see Mac Lane and Moerdijk, 1992, VI.5). Moreover, the logical connectives are interpreted as operations on Ω. This account of interpreting logic in toposes is rather different in appearance from the one we have described, though for coherent logic they are equivalent.

An elementary topos need not have a *natural numbers object* (characterized as initial induction algebra). However, it is of vital importance when it is present since then analogues of Theorems 8.25 and 8.26 will hold. Johnstone, 2002b, D5.3.5 covers the case of free algebras over sets for finitarily presented single-sorted algebraic theories.

4.4 Classifying toposes

We can now give the definition of classifying topos, as Lindenbaum algebra for predicate geometric theory. This is the analogue of Proposition 8.3, though note that we have replaced homomorphisms by "maps", going in the opposite direction.

Suppose T is a geometric type theory, X a Grothendieck topos and M a model of T in $\mathcal{S}X$. Then for every Grothendieck topos W we have a functor

$$(-)^*(M) : \mathrm{Map}(W, X) \to \mathrm{Mod}_{\mathcal{S}W}(T)$$

that takes a map f to the model $f^*(M)$. This is indeed a model, because f^* preserves all the geometric structure used to define modelhood.

Note also that it is a *functor*—natural transformations between maps are taken to homomorphisms between models (exercise!).

DEFINITION 8.39 *Let T be a geometric type theory. A* classifying topos *for T is a Grothendieck topos $[T]$, equipped with a* generic *model G of T in $\mathcal{S}[T]$, such that for every Grothendieck topos W the functor $(-)^*(G)$ is an equivalence of categories.*

We adapt the definition from Johnstone, 2002a, B4.2.1(b) with changes of notation.

Note the effect of having only an equivalence. The correspondence between models and maps is only up to isomorphism—if M is a model in $\mathcal{S}W$, then there is some map $f : W \to [T]$ such that $M \cong f^*(G)$. However, given maps f and g, there is a bijection between homomorphisms $f^*(G) \to g^*(G)$ and natural transformations $f \Rightarrow g$. A consequence of this is that the classifying topos itself is defined only up to categorical equivalence.

We now have an alternative reading to the symbol $\mathcal{S}$. If $\mathcal{S}$ stands for "sets", then $\mathcal{S}[T]$ can be read as "the category of sets with a model of T freely adjoined". This is in line with some existing notation (Johnstone, 2002a, B4.2.1) and is analogous to notation such as $\mathbb{R}[X]$ for a polynomial ring.

There are some crucial results that cannot be proved without a closer examination of the structure of classifying toposes. We defer that to Sec. 5, and meanwhile look at the use of classifying toposes. The crucial results are:

- Every geometric theory has a classifying topos (Theorem 8.65).
- Every geometric type theory has a classifying topos (Theorem 8.66, with some restrictions on the generality).
- Every Grothendieck topos classifies some geometric theory (Theorem 8.67).
- For propositional geometric theories, the maps between the locales are equivalent to the maps between their classifying toposes (Theorem 8.71). Hence for these the locale and topos treatments are equivalent, and locales can be considered a special case of toposes.

Just as for locales, we define a *point* of a topos X at *stage* W to be a map from W to X. Then maps act as point transformers by postcomposition.

The empty theory $(\emptyset, \emptyset)$ with no symbols and no axioms has a unique model in any category, given by the vacuous interpretation, so $\mathrm{Mod}_{\mathcal{S}W}(\emptyset, \emptyset)$ is the category with one object and one (identity) morphism. We write 1 for its classifying topos.

PROPOSITION 8.40 $\mathcal{S}1 \simeq \mathbf{Set}$.

Proof Mac Lane and Moerdijk, 1992, Sec. VII.1. Every set A is a coproduct of copies of the terminal object (singleton set), which we shall also write 1. (There should be no confusion between the different 1s.) Hence for any Grothendieck topos X an inverse image functor $!^* : \mathbf{Set} \to \mathcal{S}X$ has to take each set A to a coproduct of an A-indexed family of copies of 1. Moreover, any such functor preserves finite limits. (This is non-trivial, and relies on the properties of Grothendieck toposes.) The category of such functors is equivalent to the category with one object and one morphism. In other words, there does exist such a functor, and for any two such functors (with different choices of coproducts) there is a unique natural isomorphism between them. It is easy to show from the adjunction that if B is an object of $\mathcal{S}X$ then $!_*(B)$ is isomorphic to the set of global elements of B, morphisms $1 \to B$. QED

As before, points at stage 1 are called *global*. The global points of $[T]$ are equivalent to the models of T in $\mathbf{Set}$.

Let $\mathbb{O}$ be the theory with one sort and no functions, predicates or axioms. Categorically, it is the theory of "objects", since a model in $\mathcal{S}X$ is just an object of $\mathcal{S}X$, and its classifying topos is called the *object classifier* (not to be confused with the subobject classifier Ω that exists in any elementary topos). In generalization of the terminology for spaces, we call the objects of $\mathcal{S}X$ the

sheaves over X, and they are equivalent to maps from X to $[\mathbb{O}]$. Since the global points of $[\mathbb{O}]$ are sets, our intuition is that $[\mathbb{O}]$ is "the space of sets", and in Sec. 5 we shall see why it is a reasonable intuition to think of a sheaf over X as a continuous map from X to a space of sets.

4.5 Maps between classifying toposes

Now that our "Lindenbaum algebras" are Grothendieck toposes, we can—as we have seen—interpret large amounts of ordinary mathematics internally in them. This makes them very different from the lattices we used for propositional logics, and this has a profound effect on the way we can use these logical techniques. It makes it possible to treat classifying toposes $[T]$ in a very spatial way.

Suppose T_1 and T_2 are two geometric theories. By definition of classifying toposes, a geometric morphism $f : [T_1] \to [T_2]$ is equivalent to a model M of T_2 in $\mathcal{S}[T_1]$. Now all the objects and morphisms in $\mathcal{S}[T_1]$ are constructed out of the generic model G of T_1, and indeed can be constructed using finite limits and arbitrary colimits. It follows that M too has to be constructed out of the generic T_1-model. Let us portray this naively as a model transformation.

1 We declare "Let G be a model of T_1."

2 We construct a model M of T_2.

Within the scope of the declaration (1), our logic and mathematics are to be interpreted in $\mathcal{S}[T_1]$ with G the generic T_1-model. This means it must be constructively valid. We thus have a temporary change of mathematics. Back outside the scope of the declaration, returning to our ambient mathematics, we find our model construction gives a geometric morphism $f : [T_1] \to [T_2]$.

The same technique also works for natural transformations. If we define *two* models M and M', and a homomorphism $\theta_G : M \to M'$, then that gives us two maps $f, f' : [T_1] \to [T_2]$ and a natural transformation $\theta : f \Rightarrow f'$.

On the face of it, in step (2) we could use any mathematics validly interpretable in $\mathcal{S}[T_1]$. For instance, we might use Ω or function types, since in fact $\mathcal{S}[T_1]$ is an elementary topos. However, there are good reasons for restricting to geometric constructions.

If we have a point x of $[T_1]$ at stage W—that is to say, a model $x^*(G)$ of T_1 in $\mathcal{S}W$, G being the generic model of T_1, then we can apply f to it by composition and get a model $x^*(M)$ of T_2 in $\mathcal{S}W$, corresponding to $f \circ x$. If the construction of M from G (in $\mathcal{S}[T_1]$) is geometric, then it is preserved by x^*, and so the same construction constructs $x^*(M)$ out of $x^*(G)$. Hence the geometric construction works uniformly, not only for the generic point but for all points.

We therefore see that geometric morphisms between classifying toposes can be viewed as *geometric* model transformations.

It is the geometric working that enables us to view a topos spatially as comprehending all generalized points, because it allows us to transport our mathematics from one stage of definition to another along inverse image functors. Since, as we shall see later, geometric morphisms generalize continuous maps in topology, another way to view the role of the geometric constructions is that they have an intrinsic continuity.

This same view of map also provides a good way to think about generalized points. A point of X at stage Y, in other words a map $Y \to X$ is conveniently thought of as a point of X "parametrized by" a variable point of Y.

EXAMPLE 8.41 Reduct maps. *Let $F : U \to T$ be a theory morphism between geometric theories (Definition 8.19). Then every model of T is trivially a model of U by model reduction. This defines a* reduct map $\mathrm{Red}_F : [T] \to [U]$.

In fact *any* geometric morphism can be expressed as a reduct map.

THEOREM 8.42 *Let $f : [T] \to [U]$ be a geometric morphism. Then there is a geometric theory T' equivalent to T and with a theory morphism $F : U \to T'$ such that f factors as $[T] \simeq [T'] \xrightarrow{\mathrm{Red}_F} [U]$.*

Proof In $\mathcal{S}[T]$ we have the generic model G of T and a model M of U given by f. The result can be proved using the conventional techniques of Sections 5.3 and 5.4. It appears in detail in Viglas, 2004. However, here is a more conceptual reason. Each ingredient of M can be constructed from the ingredients of G using colimits and finite limits. T can be extended with sorts for such colimits or finite limits, together with structure and axioms to force then to be those colimits or limits. The extended theory is equivalent to T, since its models are determined up to unique isomorphism by their T-reducts. But there is also an obvious theory morphism from U. QED

From this one can easily deduce results such as the following.

PROPOSITION 8.43 *Let $f_i : Y_i \to X$ be a map between Grothendieck toposes ($i = 1, 2$). Then there is a pseudo-pullback square*

$$\begin{array}{ccc} Z & \xrightarrow{p_1} & Y_1 \\ p_2 \downarrow & \cong & \downarrow f_1 \\ Y_2 & \xrightarrow[f_2]{} & X \end{array}$$

(A pseudo-*pullback is like a pullback except that the square is required to commute only up to isomorphism.)*

Proof Suppose Y_i and X classify theories T_i and U. We can factor each f_i as an equivalence followed by a reduct map for a theory morphism $F_i : U \to T_i'$. Hence, we may assume that each f_i is already a reduct map. Now define P as follows. It has T_1 and T_2 put together disjointly, giving a theory morphism from each T_i. This now has two copies of U, the images of the two theory morphisms. Add function symbols and axioms to make mutually inverse homomorphisms between those two copies of U. Then a model of P comprises a model M_i of each T_i, and an isomorphism between their U-reducts, and this is exactly what is needed for the pseudo-pullback property for $Z = [P]$. QED

Hence, the points of Z are equivalent to triples (y_1, y_2, θ) where each y_i is a point of Y_i and $\theta : f_1(y_1) \cong f_2(y_2)$ is an isomorphism. When $X = 1$, we get a *product* $Y_1 \times Y_2$ whose points are pairs of points from Y_1 and Y_2.

By similar means we can construct a topos Z' whose points (y_1, y_2, θ) have $\theta : f_1(y_1) \to f_2(y_2)$ a homomorphism. The resulting square

$$\begin{array}{ccc} Z' & \xrightarrow{p_1'} & Y_1 \\ p_2' \downarrow & \overset{\theta}{\Leftarrow} & \downarrow f_1 \\ Y_2 & \underset{f_2}{\longrightarrow} & X \end{array}$$

is called a *comma square*. It does not commute, but has a natural transformation from $f_1 \circ p_1'$ to $f_2 \circ p_2'$.

Our next result shows vividly how geometric morphisms between classifying toposes can appear like functors between model categories—indeed, by taking points one extracts the functors. But as a geometric morphism, it carries extra information that it has continuity properties—for example, that it preserves filtered colimits. (This fact about geometric morphisms was exploited in Viglas, 2004 for proving that certain functors preserved filtered colimits.)

Its notion of adjunction between toposes is technically possible because the natural transformations make the category of toposes and maps into a 2-category. An adjunction between X and Y comprises maps $F : X \to Y$ and $G : Y \to X$ (the left and right adjoints), and natural transformations $\eta : \mathrm{Id}_X \Rightarrow G \circ F$ and $\varepsilon : F \circ G \Rightarrow \mathrm{Id}_Y$ such that the two composites

$$(F \circ \eta); (\varepsilon \circ F) : F \Rightarrow F \circ G \circ F \Rightarrow F$$
$$(\eta \circ G); (G \circ \varepsilon) : G \Rightarrow G \circ F \circ G \Rightarrow G$$

are both identities (cf. Mac Lane, 1971, IV.1 Theorem 2(v)).

This can all be worked through in terms of geometric transformations. However, Viglas, 2004 simplifies it greatly. Once the maps F and G have been defined, it suffices to use a geometric argument of the following form (where x and y are points of X and Y respectively). It is analogous to a more familiar characterization of adjunction, but the geometricity guarantees all the functoriality and naturality required.

- For each specialization (homomorphism) $\phi : x \Rightarrow G(y)$, define $\alpha(\phi) : F(x) \Rightarrow y$.
- For each $\psi : F(x) \Rightarrow y$, define $\beta(\psi) : x \Rightarrow G(y)$.
- Show $\beta(\alpha(\phi)) = \phi$ and $\alpha(\beta(\psi)) = \psi$.

THEOREM 8.44 *Let $F : T_1 \to T_2$ be a morphism between two Cartesian theories. Then the reduct map* $\mathrm{Red}_F : [T_2] \to [T_1]$ *has a left adjoint* $\mathrm{Free}_F : [T_1] \to [T_2]$.

Proof Constructing a free T_2-model over a T_1-model is geometric, and so the map Free_F is defined by saying for any T_1-model M, $\mathrm{Free}_F(M)$ is the free T_2-model over it.

The adjunction arises here because the corresponding adjunction within any Grothendieck topos is geometric. (This can be proved from Proposition 8.33.)

QED

4.6 Localic toposes

We now have two ways to deal with propositional geometric theories: as locales or as toposes (which in this case are called *localic*). Theorem 8.71 will show that locales and localic toposes are equivalent, but for the moment we look at some of the topos behaviour in its own right.

DEFINITION 8.45 *A topos is* localic *if it classifies a propositional geometric theory.*

A geometric type theory is essentially propositional *if it has no sorts.*

The theory Ded of Dedekind sections (Sec. 3.5) is essentially propositional.

We conjecture that the next result holds more generally, for instance when type constructs are applied to propositions (as subsingletons). However, our restricted proof is at least enough to cover our applications.

THEOREM 8.46 *Let T be an essentially propositional geometric type theory whose types can all be constructed in the empty theory. Then $[T]$ is localic.*

Proof If τ is a type in T, it has an interpretation $[|\tau|]_X$ in any Grothendieck topos $\mathcal{S}X$. We write $[|\tau|]$ for $[|\tau|]_1$ (in $\mathbf{Set}$).

We may assume without loss of generality that T is presented without any function symbols, but only predicates. This is because any function symbol can be replaced by a predicate for its graph, with axioms for single-valuedness and totality. We now show how T may be converted into an equivalent propositional geometric theory T'.

The propositions of T' are as follows. If $S \subseteq \tau_1 \times \ldots \times \tau_n$ is a predicate symbol in T, then for each $\vec{a} \in \prod_{i=1}^{n}[|\tau_i|]$ we introduce a proposition $\bar{S}_{\vec{a}}$. Now for each formula in context $(\vec{x}.\phi)$ in T, and for each $\vec{a} \in \prod_i[|\sigma(x_i)|]$, we define a formula $\bar{\phi}_{\vec{a}}$ in T' by induction as follows.

1 $(\overline{\bigvee_j \phi_j})_{\vec{a}} = \bigvee_j (\overline{\phi_j})_{\vec{a}}$, and similarly for conjunctions.

2 $\overline{(x_i = x_j)}_{\vec{a}} = \bigvee\{\top \mid a_i = a_j\}$.

3 $(\overline{(\exists y)\ \phi})_{\vec{a}} = \bigvee\{\bar{\phi}_{\vec{a},b} \mid b \in [|\sigma(y)|]\}$.

Finally, for each axiom $(\forall \vec{x})\ (\phi \longrightarrow \psi)$ of T, we give T' axioms $\bar{\phi}_{\vec{a}} \longrightarrow \bar{\psi}_{\vec{a}}$ $(\vec{a} \in \prod_i[|\sigma(x_i)|])$.

Note that, even if T is a *coherent* type theory (no infinitary disjunctions) and *finitely presented* (only finitely many symbols and axioms), T' is likely to have infinitely many symbols and axioms, and infinitary disjunctions.

We now show that T and T' are equivalent. In $\mathbf{Set}$, $[|\tau|] \cong \sum_{a \in [|\tau|]} 1$, and since this is geometric it also holds in any $\mathcal{S}X$. It follows that subobjects of $[|\tau|]_X$ correspond to $[|\tau|]$-indexed families of subobjects of 1 in $\mathcal{S}X$. Hence structures for T are equivalent to structures for T'. Now suppose that M is a structure for T, and M' the corresponding structure for T'. By structural induction on the formula ϕ, one can then show that for any formula in context $(\vec{x}.\phi)$, the subobject $\{M|\phi\}$ of $\prod_i\{M|\sigma(x_i)\}$ corresponds to the family of subobjects $\{M'|\bar{\phi}_{\vec{a}}\}$ for $\vec{a} \in \prod_i[|\sigma(x_i)|]$. From this one deduces that M is a model for T iff M' is a model for T'. QED

PROPOSITION 8.47 *Let X be a localic topos and let x and x' be points of it. Then there is at most one homomorphism from x to x'.*

Proof We can take $X = [T]$ where T is propositional. But then with no sorts, a homomorphism needs to supply no carrier functions. The sole requirement is that if P is a propositional symbol and $\{x|P\}$ holds (topologically, x is in the open P), then so does $\{x'|P\}$. QED

Hence if X is localic then $\mathrm{Map}(Y, X)$ is a preorder. We write $x \sqsubseteq x'$ if there is a homomorphism from x to x'; this is called the *specialization order* on points. Later we shall prove that this agrees with the specialization order we have already defined for locales.

Sec. 4.5 showed that maps between toposes can be defined as geometric model transformations, and this still applies to localic toposes. But Theorem 8.71 will show that those also define maps between the locales. If the locales are spatial, Proposition 8.16 shows that we then get continuous maps between the spaces. Thus geometricity of the model transformation is enough to guarantee continuity. As a logical approach to continuity, geometric logic

works by starting with ordinary logic and then *removing* the structure (e.g. negation) that makes it possible to define non-continuous functions. Compare this with other approaches, such as topology itself, or the modal logic of interior, that work by *adding* structure to support a bureaucracy of continuity proofs.

4.7 Example: the reals

We have now seen two geometric theories that purport to represent the real line. In Sec. 2.5, $T_{\mathbb{R}}$ was a propositional geometric theory described as the localic reals, while in Sec. 3.5 Ded was a predicate geometric theory whose models are the Dedekind sections. We now show that they are equivalent.

By the proof of Theorem 8.46 we see that $T_{\mathbb{R}}$ is equivalent to a theory $T'_{\mathbb{R}}$ with a single predicate symbol $P \subseteq \mathbb{Q}^2$, and axioms

$$P(q,r) \wedge P(q',r') \vdash\dashv_{qrq'r'} (\exists st)(P(s,t) \wedge \max(q,q') < s < t < \min(r,r'))$$
$$0 < \varepsilon \vdash_{\varepsilon} (\exists q)P(q-\varepsilon, q+\varepsilon)$$

Given a model of $T'_{\mathbb{R}}$, we define a Dedekind section (L,R) geometrically by

$$L = \{q \in \mathbb{Q} \mid (\exists r)P(q,r)\}$$
$$R = \{r \in \mathbb{Q} \mid (\exists q)P(q,r)\}.$$

It is easy to see that this is a Dedekind section. If $q < r$, let $\varepsilon = (r-q)/2$ and find s such that $P(s-\varepsilon, s+\varepsilon)$. If $q > s-\varepsilon$ and $r < s+\varepsilon$ then $q+\varepsilon > s > r-\varepsilon$ and $r-q < 2\varepsilon$, a contradiction. Hence either $q \leq s-\varepsilon \in L$ or $r \geq s+\varepsilon \in R$. (Note that the order on $\mathbb{Q}$ is decidable, so we can use this proof by contradiction.) Also, $P(q,r)$ holds iff $q \in L$ and $r \in R$.

Conversely, suppose (L,R) is a Dedekind section, and define $P(q,r)$ if $q \in L$ and $r \in R$. The first axiom for P is clearly satisfied. For the second, take $\varepsilon > 0$ and find $q_0 \in L$ and $r_0 \in R$ so that $P(q_0, r_0)$. Find $n \in \mathbb{N}$ such that $r_0 - q_0 < 2^{n+1}\varepsilon$. By induction on n, we show that there is some u with $P(u-\varepsilon, u+\varepsilon)$. If $n = 0$, we can take $u = (q_0+r_0)/2$, for $u-\varepsilon < q_0 < r_0 < u+\varepsilon$. Now suppose $n \geq 1$. Let $s_i = q_0 + i(r_0-q_0)/4$ $(0 \leq i \leq 4)$. Since $s_1 < s_2 < s_3$, we have (1) either $s_1 \in L$ or $s_2 \in R$, and (2) either $s_2 \in L$ or $s_3 \in R$. Examining the possibilities, we can find q_1 and r_1 from amongst the s_is with $P(q_1, r_1)$ and $r_1 - q_1 = (r_0-q_0)/2 < 2^n\varepsilon$.

The above geometric constructions give us maps $f : [T'_{\mathbb{R}}] \to [\,\mathrm{Ded}]$ and $g : [\,\mathrm{Ded}] \to [T'_{\mathbb{R}}]$. Composing them, we see that $g \circ f \cong \mathrm{Id}_{[T'_{\mathbb{R}}]}$ and $f \circ g \cong \mathrm{Id}_{[\,\mathrm{Ded}]}$. Hence the two theories are equivalent.

We have proved this solely on the hypothesis that the classifying toposes exist. We have not had to analyse the structure of the classifying toposes at all, beyond the knowledge that they are Grothendieck toposes and have generic models.

Having shown these theories are equivalent, it is possible now to *define* the real line $\mathbb{R}$ to be the classifying topos $[\,\mathrm{Ded}]$, the "space of Dedekind sections". This may seem heavy-handed. However, it tells us what the real numbers are (the points of $\mathbb{R}$, i.e. the models of Ded). It also defines the topology. The opens of $[\,\mathrm{Ded}]$ (i.e. the subobjects of 1 in $\mathcal{S}[\,\mathrm{Ded}]$—see Sec. 4.4) are equivalent to those of $[T_{\mathbb{R}}]$, and Theorem 8.71 will show that they are equivalent to elements of the frame $\Omega[T_{\mathbb{R}}]$ as defined in Sec. 2.5.

Let us use this definition of $\mathbb{R}$ to define a map.

EXAMPLE 8.48 *Addition* $+ : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is defined as follows. If* x *and* y *are points of* $\mathbb{R}$*, then* $q < x + y$ *if* $q = q_1 + q_2$ *for some* $q_1 < x$ *and* $q_2 < y$*, and* $x + y < r$ *if* $r = r_1 + r_2$ *for some* $x < r_1$ *and* $y < r_2$*.*

See how we have defined a map (a geometric morphism) just by defining a geometric construction on points. Remarkably, this is enough to guarantee continuity of the corresponding function between spaces.

5. Sheaves as predicates

We have already defined sheaves as the objects of classifying toposes. This broadens the normal usage, which defines sheaves in a more technical way. In this section we analyse more closely the structure of classifying toposes and come to the technical definition. This will enable us to prove the crucial results announced in Sec. 4.4.

Sheaves were defined first over topological spaces. Grothendieck subsequently generalized the definition to sheaves over a *site*, but as he stressed, the category of sheaves (the Grothendieck topos) is more important than the site (which is in effect a particular form of geometric theory). The reason for this is essentially our "Lindenbaum algebra" methodology—it is in terms of the topos that we get a good definition of map.

In propositional logic each proposition ϕ corresponds to a map $|\phi\}$ from models to truth values. In the geometric context this corresponds to a map from the locale to $\mathbb{S}$, and in ordinary topology this is an open.

In predicate logic, each formula $(\vec{x}.\phi)$ corresponds to a function $|\vec{x}.\phi\}$ from models to sets. In a geometric context this is a map from the classifying topos to the object classifier $[\mathbb{O}]$, and in ordinary topology it is *sheaves* that provide the corresponding "continuous set-valued map".

5.1 Sheaves over a topological space

Suppose X is a space and S is a "continuous set-valued function" on it. For this to make sense, we certainly need a set $S(x)$ for each point x; this is called the *stalk* of S at x. If we let Y be the disjoint union of the stalks we shall have a projection $\pi : Y \to X$, and the stalks $S(x)$ are recoverable as fibres $\pi^{-1}(\{x\})$.

We shall introduce a class of continuous maps π, the *local homeomorphisms*, for which it turns out that sheaves can be derived by using the fibres as the stalks.

Each stalk is to be a *set*, and we take from that that the fibre $\pi^{-1}(\{x\})$, as a subspace of Y, should have the *discrete* topology. Thus within fibres, the topology of Y should be discrete. On the other hand, across the fibres we might argue that the topology of Y should be no finer than is got from X in order that the dependence of $\pi^{-1}(\{x\})$ on x should be "continuous".

Those are vague, but let us suggest that the second "across fibre" condition requires π to be an open map—it takes open sets to open sets.

For the first "within fibre" condition, consider that a space Z is discrete iff the diagonal inclusion $\Delta : Z \hookrightarrow Z \times Z$, $\Delta(z) = (z, z)$, is open. To see this, take $z \in Z$. By definition of the product topology, there are open neighbourhoods U and V of z such that $U \times V \subseteq \Delta(Z)$. By replacing U and V by their intersection, we might as well assume $U = V$. Now if $z' \in U$ then $(z, z') \in U \times U \subseteq \Delta(Z)$, so $z' = z$. Hence $U = \{z\}$ and so $\{z\}$ is open, so the topology is discrete.

Generalizing to $\pi : Y \to X$, we use a "fibrewise discrete" property that the inclusion $\Delta : Y \hookrightarrow Y \times_X Y$ should be open. ($Y \times_X Y$ is the *fibred product*, or *pullback*, $\{(y_1, y_2) \mid \pi(y_1) = \pi(y_2)\}$.) This is more than enough to imply that every fibre is discrete in its subspace topology.

PROPOSITION 8.49 *Let $\pi : Y \to X$ be a map of spaces. Then the following conditions are equivalent.*

1. *π and the diagonal inclusion $\Delta : Y \hookrightarrow Y \times_X Y$ are both open.*
2. *Each $y \in Y$ has an open neighbourhood V such that π restricted to V is a homeomorphism onto an open neighbourhood of $\pi(y)$.*

Proof First, note that $\Delta(Y)$ is open in $Y \times_X Y$ (as subspace of $Y \times Y$) iff every $(y, y) \in \Delta(Y)$ has a basic open neighbourhood in $Y \times_X Y$ that is contained in $\Delta(Y)$, in other words we can find neighbourhoods V_1 and V_2 of y such that $V_1 \times V_2 \cap Y \times_X Y \subseteq \Delta(Y)$. By restricting to $V_1 \cap V_2$ we might as well assume $V_1 = V_2$. The condition $V \times V \cap Y \times_X Y \subseteq \Delta(Y)$ says that π is 1-1 on V. To summarize, Δ is open iff every $y \in Y$ has an open neighbourhood V on which π is 1-1.

(1)⇒(2): If $y \in Y$, choose V as above. π is a continuous bijection from V onto $\pi(V)$, and since π is open, we deduce that this bijection is a homeomorphism.

(2)⇒(1): Let W be open in Y. If $y \in W$, then we can find V_y in condition (2) and $\pi(W \cap V_y)$ is open. $\pi(W)$ is the union of these open sets $\pi(W \cap V_y)$ and hence is open. Hence π is an open map. Openness of Δ follows from what we have already said. QED

DEFINITION 8.50 *Let $\pi : Y \to X$ be a continuous map between two topological spaces. π is a* local homeomorphism *(* over X*) if it satisfies the equivalent conditions of the proposition.*

If $\pi_i : Y_i \to X$ are two local homeomorphisms over X, then a morphism from π_1 to π_2 is a map $f : Y_1 \to Y_2$ such that $\pi_2 \circ f = \pi_1$. We obtain a category $\mathbf{LocHom}_X$ of local homeomorphisms over X.

It will turn out from a long train of argument that $\mathbf{LocHom}_X$ is a Grothendieck topos. Local homeomorphisms are equivalent to sheaves as presheaves, and then from the more general topos theory they are classifying toposes. However, it is an illuminating exercise to prove it directly. The geometric constructions needed in $\mathbf{LocHom}_X$ can all be constructed stalkwise by elementary means.

For any map $\pi : Y \to X$, a *local section* of π is a map $\sigma : U \to Y$, with U open in X, such that $\pi \circ \sigma = \mathrm{Id}_U$. An open V as in Proposition 8.49 (2) is equivalent to a local section of π whose image is open. The other main definition of sheaf uses sections, through the notion of *presheaf*: a presheaf on any category $\mathcal{C}$ is a contravariant functor from $\mathcal{C}$ to the category $\mathbf{Set}$ of sets. For a topological space, a presheaf on X is defined to be a presheaf on ΩX. As with any poset, the objects of ΩX are its elements, and the morphisms are the pairs (U, V) with $U \subseteq V$ – in other words, there is a single morphism from U to V provided $U \subseteq V$. A presheaf F on ΩX has a set $F(U)$ for each $U \in \Omega X$, and if $U \subseteq V$ there is a *restriction* from $F(V)$ to $F(U)$, which we shall normally write $\sigma \mapsto \sigma|U$.

A morphism of presheaves is just a natural transformation. For presheaves over a space X, this means that a morphism from F to G has a family of functions $f_U : F(U) \to G(U)$ $(U \in \Omega X)$ that commute with the restriction maps.

DEFINITION 8.51 *Let X be a topological space. A presheaf F on X is a* sheaf *if it satisfies the following* pasting *condition.*

Let $U_i \in \Omega X$ $(i \in I)$, and suppose for each i we have $\sigma_i \in F(U_i)$ such that for all i, j we have $\sigma_i|(U_i \cap U_j) = \sigma_j|(U_i \cap U_j)$. Then there is a unique $\sigma \in F(\bigcup_i U_i)$ such that for all i we have $\sigma|U_i = \sigma_i$.

A morphism of sheaves is just a presheaf morphism. We get a category $\mathcal{S}X$ of sheaves over X.

Note the uniqueness. As an immediate consequence, by taking $I = \emptyset$ we see that if F is a sheaf then $F(\emptyset)$ is a singleton. Note also that the same definition of presheaf and sheaf work over a locale.

EXAMPLE 8.52 *Let $\pi : Y \to X$ be a local homeomorphism, and let the presheaf* $\mathrm{Sect}(\pi)$ *be defined by*

$$\mathrm{Sect}(\pi)(U) = \{\sigma : U \to Y \mid \sigma \textit{ is a local section of } \pi\}.$$

The restriction maps are ordinary domain restriction of functions. Then $\mathrm{Sect}(\pi)$ *is a sheaf.*

The process is functorial. If $\pi_i : Y_i \to X$ *(*$i = 1, 2$*) are two local homeomorphisms over* X*, and* $f : Y_1 \to Y_2$ *is a morphism between them, then composition with* f *gives a sheaf morphism* $\mathrm{Sect}(\pi_1) \to \mathrm{Sect}(\pi_2)$*. We get a functor* $\mathrm{Sect} : \mathbf{LocHom}_X \to \mathcal{S}X$*.*

THEOREM 8.53 $\mathrm{Sect} : \mathbf{LocHom}_X \to \mathcal{S}X$ *is an equivalence of categories.*

Proof (Sketch) It is necessary to show that the functor Sect is full and faithful, and essentially surjective. From $\mathrm{Sect}(\pi)$ we can recover the stalks, since the stalk at x is the colimit of the sets $\mathrm{Sect}(\pi)(U)$ as U ranges over the open neighbourhoods of x. Furthermore, we can recover the topology since the images of the local sections form a base. Starting from an arbitrary sheaf F, the same construction yields a local homeomorphism whose sheaf of sections is isomorphic to F—this proves essential surjectivity.

Faithfulness is easy, but for fullness one must show that if $\pi_i : Y_i \to X$ $(i = 1, 2)$ then every sheaf morphism $\alpha : \mathrm{Sect}(\pi_1) \to \mathrm{Sect}(\pi_2)$ comes from a morphism f from π_1 to π_2. If $y \in Y_1$, find a section $\sigma : U \to Y_1$ whose image contains y. Then $f(y)$ is defined as $\alpha_U(\sigma)(\pi_1(y))$. One must prove that this definition is independent of choice of σ, that f is continuous, that $\pi_2 \circ f = \pi_1$ and that $\mathrm{Sect}(f) = \alpha$. QED

5.2 Sheaves and local homeomorphisms for toposes

For any topos X, the sheaves over X (the objects of $\mathcal{S}X$) are equivalent to the maps $X \to [\mathbb{O}]$. Hence, by the methods of Sec. 4.5, to define a sheaf S we declare "let x be a point of X" and then, geometrically, define a set $S(x)$. We therefore think of S as a continuous set-valued map on X. (However, except on global points, these are not sets in the sense of set theory, with the structure all defined through the $\in$ relation. Geometric type theory is not done that way.) We call $S(x)$ the *stalk* of S at x, and this notation also suggests we might view the sheaf as a set parametrized by a variable point of X.

Both the point transformation and the parametrization involve a radically new notion of continuity, since $[\mathbb{O}]$ has far too few opens to be a useful topological space in anything like the conventional sense. An open of $[\mathbb{O}]$, a map $[\mathbb{O}] \to \mathbb{S}$, is a geometric definition of a truth value for each set S. There are three obvious ways to do this: constant $\top$, constant $\bot$ and by the formula $(\exists a \in S)\ \top$. In effect, we have three open subspaces of "the space of sets": the whole space, the empty space, and the space of inhabited sets. We shall later (Example 8.74) be able to prove that—at least classically—these are the only three, and from the localic point of view $[\mathbb{O}]$ cannot be distinguished from $\mathbb{S}$. (Technically, $\mathbb{S}$ is the "localic reflection" of $[\mathbb{O}]$—Definition 8.72.)

EXAMPLE 8.54 *Let T be a geometric theory. The functors and natural transformations $|B\}$, $|\vec{x}.\phi\}$, $|\vec{x}.t\}$, etc. of Remark 8.22 define sheaves and sheaf morphisms.*

Stalks can be gathered together to make a new topos, analogous to a local homeomorphism. Let $\mathbb{O}, elt$ be the theory with one sort and one constant symbol, and let $p : [\mathbb{O}, elt] \to [\mathbb{O}]$ be the obvious reduct map (which forgets the constant). The points of $[\mathbb{O}, elt]$ are pairs (A, a) where A is a set (or, in general, a sheaf over a topos) and a a global element of A.

PROPOSITION 8.55 *Let X be a topos and $A : X \to [\mathbb{O}]$ a sheaf. Let X/A be the topos given by the pseudo-pullback*

$$\begin{array}{rcl} X/A & \longrightarrow & [\mathbb{O}, elt] \\ A^*p \downarrow & \cong & \downarrow p \\ X & \underset{A}{\longrightarrow} & [\mathbb{O}] \end{array}$$

Its points are pairs (x, a) with x a point of X and $a \in A(x)$. Then $\mathcal{S}(X/A)$ is equivalent to the slice *category $(\mathcal{S}X)/A$, whose objects are morphisms in $\mathcal{S}X$ with codomain A.*

Proof $\mathcal{S}(X/A)$ is got from $\mathcal{S}X$ by freely adjoining a global element $e : 1 \to A$. From this one can construct, for any $\nu : C \to A$ in $\mathcal{S}X$, the pullback along e giving an object e^*C, and the result says in effect that every object of $\mathcal{S}(X/A)$ comes from some ν in this way.

It is straightforward to check that $(\mathcal{S}X)/A$ is a Grothendieck topos. (The corresponding fact for elementary toposes is the "Fundamental Theorem of Topos Theory", Johnstone, 2002a, A2.3.) We have a functor $q^* : \mathcal{S}X \to \mathcal{S}X/A$, with $q^*(B)$ the projection $B \times A \to A$, and it preserves colimits and finite limits. In $\mathcal{S}X/A$ the final object is $\mathrm{Id}_A : A \to A$, and q^*A has a global element e given by the diagonal $\Delta : A \to A \times A$. Every object $\nu : C \to A$ is the pullback of $q^*(\nu)$ against e.

Now suppose we have a map $f : Y \to X$ with a global element $e' : 1 \to f^*(A)$. The result amounts to showing that f^* factors via q^* and a functor $r^* : \mathcal{S}X/A \to \mathcal{S}Y$ that preserves colimits and finite limits, and takes e to e'. Clearly $r^*(\nu)$ has to be (up to isomorphism) defined as the pullback of $f^*(\nu)$ against e', after which it remains only to check that it has the required properties. QED

Note a corollary to this. A map $X/B \to X/A$ over X is equivalent to a global element of $(A \times B \to B)$ in $\mathcal{S}X/B$, and this is just a morphism $B \to A$ in $\mathcal{S}X$. Hence sheaf morphisms are equivalent to maps between the corresponding fibred spaces.

In the pseudo-pullback square in the proof of Proposition 8.55, the map A^*p on the left maps (x, a) to x. The maps that arise in this way from sheaves over X are called *local homeomorphisms* or *étale* maps (see Johnstone, 2002b, C3.3.4). For each point x, the pseudo-pullback of A^*p against x is in effect the stalk at x. In fact, we have three equivalent categories to represent sheaves over X: $\mathcal{S}X$, $\mathrm{Map}(X, [\mathbb{O}])$ and the category of local homeomorphisms with codomain X. In Joyal and Tierney, 1984 the local homeomorphisms are characterized in a way analogous to the first condition of Proposition 8.49, using a topos notion of open map.

5.3 Sites

A canonical form of geometric theory is that deriving from a *site*. We give the definition from Johnstone, 2002a, A2.1.9.

DEFINITION 8.56 *Let $\mathcal{C}$ be a small category. A* coverage *J on $\mathcal{C}$ assigns to each object A of $\mathcal{C}$ a collection $J(A)$ of families $(f_i : A_i \to A \mid i \in I)$ of morphisms targeted at A, subject to the condition that for each such family in $J(A)$, and for each morphism $g : B \to A$, there is a family $(h_{i'} : B_{i'} \to B \mid i' \in I')$ in $J(B)$ such that each $g \circ h_{i'}$ factors via some f_i.*

A category equipped with a coverage is called a site.

The definition in Mac Lane and Moerdijk, 1992, III.2 Definition 1 is slightly different, as a category equipped with a *Grothendieck topology*. In this, the covering families are all required to be *sieves*, i.e. closed under precomposition. The difference is explained in Johnstone, 2002b, C2.1.8 (where a Grothendieck topology is called a *Grothendieck coverage*). Any coverage generates a Grothendieck topology that is equivalent to it for its intended purposes.

DEFINITION 8.57 *Let $(\mathcal{C}, J)$ be a site. Then the geometric theory* $\mathrm{CtsFlat}$ *$(\mathcal{C}, J)$ of* continuous flat functors over $(\mathcal{C}, J)$ *has sorts X_A and functions $u_f : X_A \to X_B$ for the objects A and morphisms $f : A \to B$ of $\mathcal{C}$, and axioms*

$(\forall x : X_A)\ u_{\mathrm{Id}_A}(x) = x$ *($A \in \mathrm{Ob}(\mathcal{C})$)*

$(\forall x : X_A)\ u_g(u_f(x)) = u_{g \circ f}(x)$ *($f : A \to B,\ g : B \to C$)*

$\bigvee_{A \in \mathrm{Ob}(\mathcal{C})} (\exists x : X_A)\ \top$

$(\forall x : X_A, y : X_B) \bigvee_{C \in \mathrm{Ob}(\mathcal{C})} \bigvee_{f : C \to A} \bigvee_{g : C \to B} (\exists z : X_C)\ (x = u_f(z) \wedge y = u_g(z))$

$(\forall x : X_A)\ (u_f(x) = u_g(x)$
$\quad \longrightarrow \bigvee_{C \in \mathrm{Ob}(\mathcal{C})} \bigvee \{(\exists z : X_C)\ x = u_h(z) \mid h : C \to A, f \circ h = g \circ h\}$
$\qquad$ *($f, g : A \to B$)*

$(\forall x : X_A) \bigvee_{i \in I} (\exists y : X_{A_i})\ x = u_{f_i}(y)$ *($(f_i : A_i \to A \mid i \in I)$ in $J(A)$)*

Its models in a Grothendieck topos $\mathcal{E}$ are the *continuous filtering functors* (or *continuous flat functors*) from $\mathcal{C}$ to $\mathcal{E}$ (Mac Lane and Moerdijk, 1992, VII Sec. 7-9). The first two axiom schemas stipulate functoriality, the next three are the flatness (or filtering property) and the final one is the continuity. Note that if $\mathcal{C}$ has all finite limits, then (Mac Lane and Moerdijk, 1992, VII.9 Corollary 3) the flat functors from $\mathcal{C}$ to a Grothendieck topos are exactly the finite limit preserving functors.

Without the final axiom in Definition 8.57 we have the theory $\mathrm{Flat}(\mathcal{C})$ of *flat functors* over $\mathcal{C}$.

EXAMPLE 8.58 *Any Cartesian theory T is equivalent to* $\mathrm{Flat}(\mathcal{C})$ *where $\mathcal{C}$ is the opposite of the category of finitely presented T-models. For a discussion of some non-Cartesian theories of the form* $\mathrm{Flat}(\mathcal{C})$*, as well as the constructive notion of "finite" used in "finitely presented" (stronger than Kuratowski finiteness) see Vickers, 2001.*

5.4 Sheaves as presheaves

In this section we return to the main question left over from Sec. 4.4: what are the sheaves over a classifying topos? It is only when this has been answered that we can be sure classifying toposes exist. We outline the proof in stages: first, theories $\mathrm{Flat}(\mathcal{C})$; then $\mathrm{CtsFlat}(\mathcal{C}, J)$; then geometric theories in general; and then geometric type theories.

LEMMA 8.59 *If $\mathcal{C}$ is a small category then the presheaf topos $\mathbf{Set}^{\mathcal{C}^{op}}$ classifies* $\mathrm{Flat}(\mathcal{C})$.

Proof (Sketch. cf. Proposition 8.40.) For any Grothendieck topos X we want a correspondence between flat functors $F : \mathcal{C} \to \mathcal{S}X$ and functors $\mathbf{Set}^{\mathcal{C}^{op}} \to \mathcal{S}X$ preserving colimits and finite limits.

The Yoneda embedding $\mathcal{Y} : \mathcal{C} \to \mathbf{Set}^{\mathcal{C}^{op}}$ acts as a *free cocompletion* of $\mathcal{C}$ (Mac Lane and Moerdijk, 1992, I.5 Corollary 4): any functor from $\mathcal{C}$ to a cocomplete category factors uniquely (up to isomorphism) via $\mathcal{Y}$ and a colimit preserving functor. This gives an equivalence between functors $\mathcal{C} \to \mathcal{S}X$ and colimit preserving functors $\mathbf{Set}^{\mathcal{C}^{op}} \to \mathcal{S}X$.

However, $\mathbf{Set}^{\mathcal{C}^{op}}$ is also a Grothendieck topos (Mac Lane and Moerdijk, 1992, I). The content of Mac Lane and Moerdijk, 1992, VII.7 Theorem 2 is then that flatness of the functor $\mathcal{C} \to \mathcal{S}X$ is equivalent to finite limit preservation by the colimit preserving functors $\mathbf{Set}^{\mathcal{C}^{op}} \to \mathcal{S}X$. (Note also that $\mathcal{Y}$ is flat.) QED

EXAMPLE 8.60 *(See Example 8.58.) Let T be a Cartesian theory, and let $\mathcal{C}$ be its category of finitely presented models so that T is equivalent to* $\mathrm{Flat}(\mathcal{C}^{op})$*. Then $\mathcal{S}[T] \simeq \mathbf{Set}^{\mathcal{C}}$. Thus, as set-valued map on points (models of T), a sheaf*

is determined by its action on the finitely presented models. This also follows from the fact that every model is a filtered colimit of finitely presented models, and maps preserve filtered colimits.

We now turn to theories $\mathrm{CtsFlat}(\mathcal{C}, J)$. Let $F : \mathcal{C} \to \mathcal{F}$ be a flat functor to a Grothendieck topos. The continuity axiom says that for each covering family $(f_i : A_i \to A)_{i \in I}$ in $J(A)$,

$$(\forall x : F(A)) \bigvee_{i \in I} (\exists y : F(A_i))\, x = F(f_i)(y).$$

Categorically, this says that the cotupled morphism $f = [F(f_i)]_{i \in I} : \sum_{i \in I} F(A_i) \to F(A)$ is epi, in other words that its image is the whole of $F(A)$.

To analyse this, we calculate what the image is in general. The result content is roughly that of Mac Lane and Moerdijk, 1992, VII.7 Lemma 2. However, we sketch a "geometric" proof that does not rely on the subobject classifier.

LEMMA 8.61 *Let $(\mathcal{C}, J)$ be a site, and let $F : \mathcal{C} \to \mathcal{S}X$ be a flat functor to a Grothendieck topos. Let $(f_i : A_i \to A)_{i \in I}$ be in $J(A)$. Then the image of*

$$f = [F(f_i)]_{i \in I} : \sum_{i \in I} F(A_i) \to F(A)$$

is the colimit of a diagram $F \circ \Delta$ as follows. Let $\mathcal{D}$ be the full subcategory of the slice category $\mathcal{C}/A$ whose objects are morphisms $g : C \to A$ that factor through some f_i. Δ is the obvious functor from $\mathcal{D}$ to $\mathcal{C}$, taking $(g : C \to A)$ to C.

Proof Let $\mu_i : F(A_i) \to \sum_i F(A_i)$ be the coproduct injection, and let $f = m \circ q$ be the image factorization of f.

We define a cocone from $F \circ \Delta$ to $\mathrm{Im}\, f$ as follows. If $g : C \to A$ factors via some f_i, then $F(g)$ factors via f and hence uniquely via m, as $m \circ \nu_g$ (say).

Now suppose we have a cocone from $F \circ \Delta$ to some K, given by morphisms $\nu'_g : F(C) \to K$ for each $g : C \to A$ in $\mathrm{Ob}(\mathcal{D})$. If we are to have a colimit morphism $\alpha : \mathrm{Im}\, f \to K$, then it is determined uniquely by $\alpha \circ q$ (because q is epi) and hence by the morphisms $\alpha \circ q \circ \mu_i = \alpha \circ \nu_{f_i} = \nu'_{f_i}$. This proves uniqueness for α.

The morphisms ν'_{f_i} give us a morphism $\alpha' = [\nu'_{f_i}]_i : \sum_i F(A_i) \to K$, and we should like α' to factor as $\alpha \circ q$ for some $\alpha : \mathrm{Im}\, f \to K$. This will be as required, since if $g = f_i \circ g' : C \to A$ then

$$\nu'_g = \nu'_{f_i} \circ F(g') = \alpha' \circ \mu_i \circ F(g') = \alpha \circ q \circ \mu_i \circ F(g') = \alpha \circ \nu_{f_i} \circ F(g') = \alpha \circ \nu_g.$$

To prove existence of α, we interpret logic in $\mathcal{S}X$. Viewing $\mathrm{Im}\, f$ as a quotient of $\sum_i F(A_i)$, it suffices to show that if $f(a) = f(b)$ $(a, b \in \sum_i F(A_i))$ then

$\alpha'(a) = \alpha'(b)$. Suppose $a = \mu_i(a')$, with $a' \in F(A_i)$, and similarly $b = \mu_j(b')$, so $F(f_i)(a') = F(f_j)(b')$. By flatness of F we can find an object C in $\mathcal{C}$, morphisms $g : C \to A_i$ and $h : C \to A_j$ with $f_i \circ g = f_j \circ h$, and $c \in F(C)$ with $a' = F(g)(c)$ and $b' = F(h)(c)$. Then

$$\begin{aligned}\alpha'(a) &= \alpha' \circ \mu_i \circ F(g)(c) = \nu'_{f_i} \circ F(g)(c) = \nu'_{f_i \circ g}(c) \\ &= \nu'_{f_j \circ h}(c) = \nu'_{f_j} \circ F(h)(c) = \alpha' \circ \mu_j \circ F(h)(c) = \alpha'(b).\end{aligned}$$

QED

At this point we can introduce the notion of *sheaf* over a site. (In Mac Lane and Moerdijk, 1992, III.4 this is for a slightly different definition of site, but the difference is not of great significance here.)

DEFINITION 8.62 *Let $(\mathcal{C}, J)$ be a site, and let S be a presheaf over $\mathcal{C}$. S is a* sheaf *if it has the following* pasting *property.*

Suppose $(f_i : A_i \to A)_{i \in I}$ is in $J(A)$. Suppose for each $i \in I$ we have $x_i \in S(A_i)$, with the family $(x_i)_{i \in I}$ "matching" in the sense that if C is an object of $\mathcal{C}$ and $g_i : C \to A_i$, $g_j : C \to A_j$ are morphisms with $f_i \circ g_i = f_j \circ g_j$ then $S(g_i)(x_i) = S(g_j)(x_j)$. Then there is a unique $x \in S(A)$ such that $x_i = S(f_i)(x)$ for all i.

LEMMA 8.63 *Suppose $F : \mathcal{C} \to \mathcal{S}X$ is flat, and let $f : X \to [\mathrm{Flat}(\mathcal{C})]$ be the corresponding map. Then F is continuous iff for every object U of $\mathcal{S}X$, the presheaf $f_*(U)$ is a sheaf.*

Proof Suppose $(f_i : A_i \to A)_{i \in I}$ is in $J(A)$. By Yoneda's Lemma, if S is a presheaf then a family $(x_i)_{i \in I}$ of elements $x_i \in S(A_i)$ corresponds to a family $(\xi_i)_{i \in I}$ of morphisms $\xi_i : \mathcal{Y}(A_i) \to S$. Let $\Delta : \mathcal{D} \to \mathcal{C}$ be the diagram described in Lemma 8.61. Then we find that the family $(x_i)_i$ is matching iff the family $(\xi_i)_i$ extends (uniquely) to a cocone from $\mathcal{Y} \circ \Delta$ to S. The existence of x is equivalent to the factorization of this cocone through $\mathcal{Y}(A)$. If S is of the form $f_*(U)$, then the cocone of presheaves corresponds to a cocone in $\mathcal{S}X$ from $f^* \circ \mathcal{Y} \circ \Delta = F \circ \Delta$ to U. By Lemma 8.61 we know F is continuous iff for every covering (f_i) we have that the colimit of $F \circ \Delta$ is $F(A)$, i.e. for every U, every cocone $F \circ \Delta \to U$ factors via $F(A) = f^* \circ \mathcal{Y}(A) \to U$ and this gives us our $\mathcal{Y}(A) \to f_*(U)$ as required for finding x. Hence F is continuous iff for every covering and for every U we can perform the pasting with $f_*(U)$. But this just says that every $f_*(U)$ is a sheaf. QED

If we define $\mathrm{Sh}(\mathcal{C}, J)$ to be the full subcategory of $\mathbf{Set}^{\mathcal{C}^{op}}$ comprising the sheaves, then we see that the maps $X \to [\mathrm{Flat}(\mathcal{C})]$ corresponding to continuous flat functors are the ones whose direct image part factors via $\mathrm{Sh}(\mathcal{C}, J)$.

THEOREM 8.64 *If* $(\mathcal{C}, J)$ *is a site then* $\mathrm{Sh}(\mathcal{C}, J)$ *is a classifying topos for* $\mathrm{CtsFlat}(\mathcal{C}, J)$.

Proof This is the content of Mac Lane and Moerdijk, 1992, VII.9 Corollary 2, which states that—in the conventional notation—there is an equivalence of categories between $\mathrm{Map}(\mathcal{E}, \mathrm{Sh}(\mathcal{C},J))$ and the category of continuous filtering functors $\mathcal{C} \to \mathcal{E}$ (i.e. models of the site theory). In our notation we can thus take $\mathrm{Sh}(\mathcal{C},J)$ as $\mathcal{S}[\mathrm{CtsFlat}(\mathcal{C}, J)]$. In outline, the rest of the proof is as follows.

First, $\mathrm{Sh}(\mathcal{C}, J)$ is indeed a topos. (This includes the fact that it is an elementary topos. This is perhaps unexpected, since the argument from classifying toposes worked with the geometric structure.)

Next, the inclusion $\mathrm{Sh}(\mathcal{C}, J) \to \mathbf{Set}^{\mathcal{C}^{op}}$ is the direct image part of a geometric morphism. Proving the existence of the inverse image part, the "associated sheaf functor" or *sheavification*, is of fundamental importance. If S is already a sheaf, then it is its own sheavification.

After all that, proving that $\mathrm{Sh}(\mathcal{C}, J)$ classifies flat continuous functors is more or less Lemma 8.63. QED

We can also calculate the stalks explicitly. Let x be a global point, a continuous flat functor from $\mathcal{C}$ to $\mathbf{Set}$, and S a sheaf. The stalk $S \circ x$ can be calculated in two stages (Mac Lane and Moerdijk, 1992, Sec. VII.5). First, let U_0 be the disjoint union over all objects A of $\mathcal{C}$ of the products $x(A) \times S(A)$. Next, if $f : A \to B$ is a morphism in $\mathcal{C}$, and $a \in x(A)$ and $b \in S(B)$, we identify $(a, S(f)(b))$ and $(x(f)(a), b)$ in U_0 and generate an equivalence relation $\sim$ thereby. Then the stalk is $U_0/{\sim}$. This construction is geometric, and can be reproduced for non-global points.

THEOREM 8.65 *Every geometric theory is equivalent to a site theory* $\mathrm{CtsFlat}$ $(\mathcal{C}, J)$*, and hence has a classifying topos.*

Proof Let T be a geometric theory over signature Σ. By Johnstone, 2002b, Lemma D1.3.8, every geometric formula in context over Σ is logically equivalent to one of the form $\bigvee_i (\exists \vec{y}_i)\phi_i$ where each ϕ_i is a Horn formula (a conjunction of equations and predicate symbols applied to terms). It follows that each axiom in T is equivalent to a set of axioms of the form $\psi \vdash_{\vec{x}} \bigvee_i (\exists \vec{y}_i)\phi_i$. Moreover, by replacing ϕ_i by $\psi \wedge \phi_i$ and using the distributivity and Frobenius rules, we may assume that $\phi_i \vdash_{\vec{x}\vec{y}_i} \psi$. From Σ can be constructed (Johnstone, 2002b, D1.4) a *syntactic category* $\mathcal{C}$, Cartesian (i.e. with all finite limits), such that in any Cartesian category $\mathcal{D}$ we have that interpretations of Σ in $\mathcal{D}$ are equivalent to Cartesian (finite limit preserving) functors from $\mathcal{C}$ to $\mathcal{D}$; and recall that because $\mathcal{C}$ is Cartesian, flat functors from $\mathcal{C}$ to a topos are the same as Cartesian functors. The objects of $\mathcal{C}$ are the Horn formulae in context, modulo renaming of variables, and the morphisms are the formulae that are "provably the graphs

of functions", modulo logical equivalence. Now suppose $\psi \vdash_{\vec{x}} \bigvee_i (\exists \vec{y}_i)\phi_i$ is one of the axioms in T. In $\mathcal{C}$ we have diagrams

$$\begin{array}{ccc} (\vec{x}, \vec{y}_i.\phi_i) & \hookrightarrow & (\vec{x}, \vec{y}_i.\top) \\ \downarrow & & \downarrow \\ (\vec{x}.\psi) & \hookrightarrow & (\vec{x}.\top) \end{array}$$

where the right-hand arrow is the product projection, and the left-hand arrow follows from our assumption that $\phi_i \vdash_{\vec{x}\vec{y}_i} \psi$. We take those left-hand arrows, as i varies, as covering $(\vec{x}.\psi)$, and use these covers to generate a coverage J of $\mathcal{C}$. Models of T are equivalent to models of $\mathrm{CtsFlat}(\mathcal{C}, J)$. QED

We should now like a result of the form "every geometric type theory has a classifying topos". This is difficult, since our notion of geometric type theory is only informal. The following argument from Johnstone, 2002a, B4.2 uses a particular restricted formalization that nonetheless seems ample to cover examples that arise in practice.

THEOREM 8.66 *Normally, geometric type theories have classifying toposes. (The proof is not completely general.)*

Proof Johnstone, 2002a, Definition B4.2.7(c) gives a definition of geometric theory that includes features of geometric type theory. According to that definition, a geometric theory T is built up in a finite sequence $T_0, \ldots, T_n = T$. T_0 declares finitely many sorts, and each subsequent step is of one of two forms. A *simple functional extension* T_{i+1} of T_i declares a function symbol $f : F_1 \to F_2$, where F_1 and F_2 are geometric types. A *simple geometric quotient* T_{i+1} of T_i is based on a morphism $u : F_1 \to F_2$ of geometric types. T_{i+1} adds axioms

$$\begin{aligned} u(x) = u(x') &\vdash_{x,x':F_1} x = x' \\ \top &\vdash_{y:F_2} (\exists x : F_1)\, y = u(x) \end{aligned}$$

and thus forces u to be an isomorphism.

In each case, if T_i has a classifying topos, then we can identify the geometric types (F_1, F_2) and morphisms (u) with objects and morphisms of $\mathcal{S}[T_i]$, and one can construct a classifying topos for T_{i+1}. Hence every geometric theory by that definition has a classifying topos.

These two steps provide a completely general way of introducing function symbols, and also axioms $\phi \vdash_{\vec{x}} \psi$, for satisfaction of the axiom is equivalent to saying that the inclusion morphism $\psi \wedge \phi \to \phi$ is an isomorphism. As for predicate symbols $P \subseteq \vec{A}$, these can be introduced with a sort P' and function $i_P : P' \to \vec{A}$ which must then be constrained to be monic (to give a subobject corresponding to P). This is done by an axiom

$$i_P(x) = i_P(x') \vdash_{x,x':P'} x = x'.$$

Hence all the ingredients of geometric type theory can be introduced by these steps. QED

Since only finitely many steps are allowed, it would seem that the geometric type theory according to that definition should be finitely presented—only finitely many symbols and axioms. However, in practice one can get round that by internalizing the indexing set of an infinite family of symbols or axioms. For example, consider modules over a ring R. The algebraic theory of these would normally be presented with a (possibly infinite) R-indexed family of unary operators σ_r for scalar multiplication. But the set R is a constant geometric type (a coproduct of an R-indexed family of copies of 1) over any theory, and modules M can equivalently be presented using an operator $\sigma : R \times M \to M$. (Exercise: formulate this using simple functional extensions and simple geometric quotients.)

THEOREM 8.67 *Let $\mathcal{E}$ be a category. Then the following are equivalent.*

1 *$\mathcal{E}$ is a Grothendieck topos (as defined in Definition 8.31).*

2 *$\mathcal{E}$ is equivalent to* $\mathrm{Sh}(\mathcal{C}, J)$ *for some site* $(\mathcal{C}, J)$.

3 *$\mathcal{E}$ is classifying topos for some geometric theory.*

Proof (1)⇔(2) is known as *Giraud's Theorem*. See Johnstone, 2002b, C2.2.8, where condition (vii) is our condition (1). For an alternative version, see Mac Lane and Moerdijk, 1992, Appendix, Theorem 1. (2) is usually taken as the definition of Grothendieck topos.

(2)⇔(3): Theorems 8.64 and 8.65. QED

5.5 Sheaves for locales

We now turn to the question of how continuous maps between spaces and locales relate to geometric morphisms between toposes.

PROPOSITION 8.68 *Let X be a Grothendieck topos. Then* $\mathrm{Sub}_{\mathcal{S}X}(1)$ *is a frame.*

Proof Johnstone, 2002b, C1.4.7. In fact $\mathrm{Sub}_{\mathcal{S}X}(S)$ is a frame for any sheaf S. QED

Now for any propositional geometric theory T, topos models in $\mathcal{S}X$ are equivalent to frame models in $\mathrm{Sub}_{\mathcal{S}X}(1)$. It follows that T and $\mathrm{Th}_{\Omega[T]}$ (Definition 8.9) are equivalent with respect to topos models.

Let A be a frame. As a poset it can also be considered a category, and we can define a coverage J on it as follows. Let $a \in A$, and let $\{b_i \mid i \in I\} \subseteq \{b \mid$

$b \leq a\}$. Then $\{b_i \mid i \in I\} \in J(a)$ if $a \leq \bigvee_{i \in I} b_i$. (Exercise: this is indeed a coverage in the sense of Definition 8.56.)

PROPOSITION 8.69 *The theories* $\mathrm{CtsFlat}(A, J)$ *and* Th_A *are equivalent.*

Proof As a category, A is Cartesian (products are meets, and equalizers are trivial). Hence, flatness of a functor is equivalent to preservation of finite limits. The top element of A must map to the terminal object 1, and all the other elements of A to subobjects of 1 (because if a functor preserves finite limits then it preserves monics, and all the morphisms in A are monic). Hence a flat functor over A is equivalent to a function $A \to \mathrm{Sub}(1)$ that preserves finite meets. Continuity then says that the function preserves arbitrary joins too. QED

If A is a frame, then for a presheaf $S : A^{op} \to \mathbf{Set}$, if $a \leq b$ in A and $x \in S(b)$, then we write $x|_a$ for $S(a \leq b)(x)$, the *restriction* of x to a.

THEOREM 8.70 *Let A be a frame, and let X be the topos* $[\mathrm{Th}_A]$.

1 *A sheaf over X is equivalent to a sheaf over the locale for A (Definition 8.51, replacing ΩX by A, and $\cap$ and $\bigcup$ by $\wedge$ and $\bigvee$).*

2 *There is an order isomorphism between* $\mathrm{Sub}_{\mathcal{S}X}(1)$*, the set of subsheaves of* 1 *over X, and A.*

Proof (1) is calculated directly from Definition 8.62 using Theorem 8.64 and Proposition 8.69. For (2), the terminal sheaf 1 is defined by $1(a) = 1$ (i.e. some singleton) for every $a \in A$. This can be calculated directly, but it also follows from the fact that the embedding $\mathcal{S}X \to \mathbf{Set}^{A^{op}}$ is a right adjoint and hence preserves all limits, and finite limits in $\mathbf{Set}^{A^{op}}$ are calculated argumentwise. Now the subsheaves of 1 are the sheaves S for which every $S(a)$ is a subsingleton.

For every $b \in A$ we have a subsheaf S_b of 1 defined by $S_b(a) = 1$ iff $a \leq b$. (In fact these make up the generic point of X in $\mathcal{S}X$.) Clearly if $S_b = S_{b'}$ then $b \leq b' \leq b$, so $b = b'$. On the other hand, suppose S is a subsheaf of 1 and let b be the join of those $a \in A$ for which S_a is inhabited. By pasting we find that $S(b)$ is inhabited, and it follows that $S = S_b$. QED

It follows that for any propositional geometric theory T we have $\mathrm{Sub}_{\mathcal{S}[T]}(1) \cong \Omega[T]$.

THEOREM 8.71 *Let T and T' be propositional theories. Then there is an equivalence between*

1 *locale maps $[T] \to [T']$, and*

2 *topos maps* $[T] \to [T']$.

Proof A topos map $[T] \to [T']$ is equivalent to a model of $\mathrm{Th}_{\Omega[T']}$ in $\mathcal{S}[T]$, i.e. a frame homomorphism $\Omega[T'] \to \mathrm{Sub}_{\mathcal{S}[T]}(1) \cong \Omega[T]$. QED

Referring back to Proposition 8.16, we see that for sober spaces, continuous maps are equivalent to geometric morphisms between the corresponding toposes. We have now justified the key fact that underlies this chapter: toposes generalize topological spaces (at least in the sober case), and geometric morphisms are the continuous maps at topos generality.

We now know that locales and localic toposes are equivalent. We write X without any bias either way, and refer concretely to the frame as ΩX and to the category of sheaves as $\mathcal{S}X$. More generally, for any Grothendieck topos X we can write ΩX for the frame $\mathrm{Sub}_{\mathcal{S}X}(1)$ without creating any ambiguity in the localic case. We call its elements *opens* of X, equivalent to maps $X \to \mathbb{S}$.

DEFINITION 8.72 *Let X be a Grothendieck topos. Then the* localic reflection *of X is the locale* $\mathrm{Loc}(X)$ *whose frame is ΩX.*

PROPOSITION 8.73 *Let X be a Grothendieck topos. Then there is a map $\alpha : X \to \mathrm{Loc}(X)$ such that any map $f : X \to Y$ with Y a locale factors uniquely (up to isomorphism) via α.*

Proof This is immediate from the fact that if Y is a locale, then geometric morphisms from X to Y are equivalent to frame homomorphisms from ΩY to $\mathrm{Sub}_{\mathcal{S}X}\, 1$. QED

If $\phi : x \Rightarrow y$ is a specialization morphism between points of X, then $\alpha(x) \sqsubseteq \alpha(y)$. Hence x and y are identified by α if there are specialization morphisms going in both directions between them. Thus the localic reflection can lose a lot of structure.

EXAMPLE 8.74 *Consider the object classifier $[\mathbb{O}]$. Classically, if A and B are two sets then there is a function from A to B unless A is inhabited and B is empty. Hence we might expect $\mathrm{Loc}([\mathbb{O}])$ to have two points for two classes of sets: inhabited, and empty. We can calculate that in fact $\mathrm{Loc}([\mathbb{O}]) \simeq \mathbb{S}$. The theory $\mathbb{O}$ is algebraic, and its category of finitely presented algebras is the category Fin of finite sets. (Constructively, this is "finite" in a strong sense, meaning isomorphic to $\{1, \ldots, n\}$ for some natural number n.) Hence $\mathcal{S}[\mathbb{O}] \simeq \mathbf{Set}^{\mathrm{Fin}}$. A sheaf $S : \mathrm{Fin} \to \mathbf{Set}$ is a subsheaf of 1—an open—iff every $S(A)$ is a subsingleton, and we find it is determined up to isomorphism by $S(0) \subseteq S(1) \subseteq 1$. It can be calculated that the frame of these is isomorphic to $\Omega\mathbb{S}$.*

Thinking of $[\mathbb{O}]$ as a generalized space, we now see how far it is from being an un*generalized space. Its opens are simply too few to characterize the generalized topological structure and we have to use sheaves instead.*

6. Summary of toposes

Let us summarize the key points of this story.

1 The usual semantics of first-order logic provides meaning in sets: sorts are sets, function symbols (and terms generally) are functions, and predicates (and formulae) are subsets of products. This tells us, for each theory, what are the *models* of that theory.

2 Categorical logic uses the same idea to provide meaning in more general categories: sorts are objects, function symbols and terms are morphisms, and predicates and formulae are subobjects of products. It tells us what the models of a theory are in more general categories.

3 The logic has to be matched to the categorical structure. The ability to interpret logical connectives, and the validity of logical axioms in an interpretation, both depend on the structure and properties of the category.

4 It is natural to form axioms in two stages as $(\forall \vec{x})(\phi \to \psi)$. Then ϕ and ψ are formulae, using connectives appropriate to the categorical structure, and the form of the axioms compares two subobjects (for ϕ and ψ) and uses minimal categorical structure.

5 The logic we are particularly interested in, *geometric logic*, is interpreted in Grothendieck toposes. However, it is only a fragment of what can be interpreted there. Its formulae use $\wedge$, $\bigvee$, $=$ and $\exists$.

6 It is related to *geometric morphisms* between toposes, in that the geometric logic is preserved by the inverse image functors of geometric morphisms.

7 To emphasize the difference between spatial and logical aspects of toposes, we use a non-standard notation with simple symbols to denote a topos "as generalized topological space", and we apply an $\mathcal{S}$ to denote the same topos "as generalized universe of sets" (in other words, the category discussed above where the logic is interpreted). Thus a geometric morphism $f : X \to Y$ comprises two functors $f^* : \mathcal{S}Y \to \mathcal{S}X$ and $f_* : \mathcal{S}X \to \mathcal{S}Y$.

8 There are type constructors that can be considered to be within the scope of geometric logic. These include free algebra constructions. Although we have not defined the precise range of these type constructors, we have introduced the phrase *geometric type theory* for theories that use those finitary constructors we know to be of this kind. They are equivalent in expressive power to geometric theories.

9 *Coherent theories* and *coherent type theories* are similar to the geometric versions but do not use infinitary disjunctions. Coherent type theories are intermediate in expressive power between coherent theories and geometric theories. It is found in practice that once the finitary type constructors are brought in, the infinitary disjunctions of geometric logic are often not needed.

10 We define a *(generalized) point* of a topos X to be a geometric morphism whose codomain is X. It is a *global* point if its domain (its *stage of definition)* is the topos 1 where $\mathcal{S}1 = \mathbf{Set}$.

11 Each geometric type theory T has a *classifying topos* $[T]$ whose points at stage Y are the models of T in $\mathcal{S}Y$. $\mathcal{S}[T]$ is generated by a "generic" model of the theory and is an analogue of Lindenbaum algebra for a predicate geometric theory.

12 The Grothendieck toposes are the classifying toposes for geometric type theories. They can be constructed as toposes of sheaves over sites.

13 A geometric morphism from X to Y transforms, by composition, points of X (at any stage of definition) to points of Y.

14 By the definition of classifying topos, we define a geometric morphism from $[T_1]$ to $[T_2]$ by constructing a model of T_2 in $\mathcal{S}[T_1]$. Since $\mathcal{S}[T_1]$ is generated by a generic model of T_1, this appears formally as declaring, "Let M be a model of T_1," and then constructing a model of T_2 out of it. To be valid in $\mathcal{S}[T_1]$, the construction must be intuitionistically valid; and to be uniform over all stages of definition it must be geometric.

15 Thus we think of Grothendieck toposes as generalized spaces of models, and geometric morphisms as maps between those spaces.

For some examples of the techniques in use, see Vickers, 1999, Vickers, 2001 and Vickers, 2004. In particular, Vickers, 2001 discusses toposes X for which $\mathcal{S}X$ is a presheaf category, with reference to examples such as the simplicial sets Mac Lane and Moerdijk, 1992, Sec. VIII.8.

7. Other directions

We have focused on the relationship between geometric logic and the categorical structure of Grothendieck toposes, to give an introduction to how toposes can be understood as generalized topological spaces. However, the connections between logic and toposes go far beyond this and most of the standard texts describe a range of broader applications. We now briefly mention just a few other aspects of topos theory that are relevant to the logic of space.

7.1 Fibred locales

We have already seen how a map $f : X \to Y$ can be understood as a generalized point of Y, continuously parametrized by a variable point of X. In terms of the non-classical mathematics of sheaves, this is a model in $\mathcal{S}X$ of whatever theory Y classifies.

However, we can also look at the parametrization the other way round. For each point y of Y, we get a fibre $X_y = f^{-1}(\{x\})$—indeed, this still makes sense for toposes, by taking the pseudo-pullback of f along y. Hence this is a space "parametrized by a variable point of Y". We have seen one example of this already, in sheaves and local homeomorphisms. There is a particular "localic" kind of map f between toposes, essentially meaning that X is presented by no new sorts relative to Y (and in particular any map between locales is localic). This gives a notion of "fibred locale" over Y, and it turns out that this is equivalent to doing locale theory constructively in $\mathcal{S}Y$.

Joyal and Tierney, 1984 give a straightforward approach to this using frames and we shall sketch that. (Vickers, 2004 gives a more geometric account.) The notion of frame (and frame homomorphism) can be defined in any elementary topos. However, the theory is not finitary algebraic and makes essential use of the elementary topos structure: to define arbitrary joins on A requires a morphism from the powerobject $\mathcal{P}(A)$ to A.

Frame structure is preserved by direct image functors f_* (though not by f^*), and the subobject classifier is Ω is always a frame. Hence for any map $f : X \to Y$, $f_*(\Omega_X)$ is a frame in $\mathcal{S}Y$. On the other hand, given a frame A in $\mathcal{S}Y$, we can replicate the construction of the category of sheaves to get a localic map $p : Z \to Y$ such that $p_*(\Omega_Z) \cong A$. In fact we find a duality between frames in $\mathcal{S}Y$ and fibred locales over Y.

EXAMPLE 8.75 *Let S be a sheaf over a topos X, and let $f : X/S \to X$ be the map of Proposition 8.55. One can calculate that the subobject classifier in $\mathcal{S}X/S$ is $S \times \Omega_X \to S$ and its image under f_* is $\mathcal{P}(S)$. Relative to X, it is therefore the discrete locale (i.e. all subsets open) corresponding to S.*

7.2 Powerlocales

Powerlocales are the localic analogue of hyperspaces, spaces whose points are subspaces of other spaces. If X is a locale, then there are various kinds of powerlocales whose points are different kinds of sublocales (the localic analogue of subspace) of X.

In some ways the starting point is the Vietoris powerlocale $\mathrm{V}\,X$, which bears a direct relationship to the Vietoris hyperspace and was first studied in Johnstone, 1985. In computer science an analogous "Plotkin powerdomain" has been used to give semantics for non-deterministic programs—that is, programs for which the result is in some sense a range of points. It was noticed (Smyth,

1978) that its topology is generated by two coarser topologies that give two powerdomains that are interesting in their own right, and these were transferred (Robinson, 1986) to locales to give the upper and lower powerlocales $\mathrm{P}_U\, X$ and $\mathrm{P}_L\, X$. Computer science applications in localic form have appeared in Abramsky, 1991a and Abramsky, 1991b. The three principal powerlocales (Vietoris, upper, lower) are summarized in Vickers, 1997. Their relationship with the predicative mathematics of formal topology is discussed in Vickers, 2006 and Vickers, 2005. More recently (Johnstone and Vickers, 1991, Vickers, 2004, Vickers and Townsend, 2004) it has been noticed that both the upper and lower powerlocales embed in a larger *double powerlocale*, which can be got as either $\mathrm{P}_U\, \mathrm{P}_L\, X$ or $\mathrm{P}_L\, \mathrm{P}_U\, X$ (they are homeomorphic).

Each powerlocale has a good logical content, long understood in computer science. Given a locale X, each powerlocale embodies a logical theory whose models are certain kinds of *sublocales* of X. A sublocale is in effect a theory got by adding extra axioms to that for X, thus specifying a part of the class of models of X. Some topological properties of X, compactness being a good example, can be discussed in terms of points of the powerlocales (Vickers, 1995, Vickers, 2006).

The logical approach relies on the idea that, given a logic of points, we get a "logic of parts", reminiscent of modal logic. For each property U of points, an open of the original locale, we get two properties of parts: $\Box U$ says that the part is wholly inside U, while $\Diamond U$ says that the part has at least one point in U (i.e. it *meets* U). Suitable axioms for the properties $\Box U$ are that $\Box$ preserves finite meets and also directed joins – this latter turns out to be necessary for good results, and imposes a compactness condition on the parts. From these we get the upper powerlocale. A suitable property for $\Diamond$ is that it preserves all joins, and from that we get the lower powerlocale. Taking the properties $\Box U$ and $\Diamond U$ together, we need extra axioms to show their interaction:

$$\Box U \wedge \Diamond V \to \Diamond(U \wedge V),$$
$$\Box(U \vee V) \to \Box U \vee \Diamond V.$$

From these we get the Vietoris powerlocale.

At that first stage, the powerlocales are defined directly in terms of the frames. However, one can also investigate them as theory constructions. That is to say, if the original space (the "logic of points") is given as a theory rather than as a frame, we show how to gain theories of the powerlocales. The proofs uses "coverage theorems", results that transform a presentation of the frame by generators and relations into a presentation of the same structure but by generators and relations with respect to different algebraic operators.

7.3 Modal logic

One direction that might particularly interest readers of this book is the connection with modal logic. The pointers that follow here were supplied by the Second Reader of this chapter. Classical as well as non-classical modalities have been studied along topos-theoretic lines by Reyes with others: see Lavendhomme et al., 1989; Reyes, 1991; Makkai and Reyes, 1995; Reyes and Zolfaghari, 1991; Reyes and Zolfaghari, 1996. Categorical semantics for super-intuitionistic and modal predicate logics were developed by Ghilardi (Ghilardi, 1989; Ghilardi, 1991; Ghilardi, 1992) and Shehtman and Skvortsov (Shehtman and Skvortsov, 1990; Skvortsov and Shehtman, 1993; Skvortsov, 1996; Skvortsov, 2003); see also Suzuki, 1990; Suzuki, 1993; Isoda, 1997; Nagaoka and Isoda, 1997; Shirasu, 1998. A modal intuitionistic calculus of nuclei was developed by Goldblatt (Goldblatt, 1981; Goldblatt, 1979).

8. Conclusions

It seems obvious, even trite, that a logic of finite conjunction and arbitrary disjunction might be related to the finite intersections and arbitrary unions of open sets in topology. Locale theory shows how propositional geometric theories can be studied topologically. Nonetheless, geometric logic is very peculiar from the perspective of traditional logic. Its incompleteness seems a grave disadvantage, while its type-theoretic content in a first-order logic comes as a surprise.

Our basic message is that in a constructive geometric mathematics, topology appears as an emergent feature: the logical theories describe classes of models with an intrinsic topology (in Grothendieck's generalized sense, using sheaves when there are not enough opens), and mathematical constructions have an intrinsic continuity.

Paradoxically, the constructivity provides the way around the incompleteness. Normally one thinks of constructivity as the enemy of completeness, because so many completeness proofs are classical. But by allowing for constructive mathematics one gains access to a more complete range of models of each geometric theory. Amongst the Grothendieck toposes each theory has its classifying topos, equipped with the generic model. It serves as "generalized Lindenbaum algebra", but can also be thought of as "the space of models". Geometric morphisms are logic-preserving functors between the toposes, but can also be used (in the reverse direction) as continuous maps of models, at any stage. This is without reference to the concrete class of standard models, of which there might anyway be insufficient because of the incompleteness. The propositional fragment can alternatively be treated using locales (and frames as Lindenbaum algebras), but the two treatments are equivalent. In the spatial

case the geometric morphisms recover the known notion of continuous map between spaces (modulo issues of sobriety).

That broad story underlies much of topos theory, though there are also many deep non-geometric uses of Grothendieck toposes.

The type theoretic content is an unfamiliar development in first order logic. In a sense it is superfluous, since it does not essentially extend the scope of geometric logic. Nonetheless it makes the logic more convenient and in particular it can be used to eliminate infinitary disjunctions in favour of finitary constructions.

Combining *coherent* logic with some of the geometric type constructors, we get a coherent type theory. This must be less expressive than geometric logic, yet it is already enough to capture important topological examples such as the real line. An exciting thought is that this may provide an example of topology emerging from a *finitary* type theory, with finite coproducts and the inductive construction of free models. Such a coherent type theory would be better described in its own terms, with a corresponding class of categories to interpret it. A promising candidate class for these categories is the arithmetic universes of Joyal. By contrast with Grothendieck toposes, these categories do not automatically have function spaces or subobject classifiers. This is going to require a much more careful syntactic formulation of the coherent type theory, probably including aspects of dependent type theory. Some preliminary results have been found in Maietti, 2003.

References

Abramsky, S. (1991a). Domain theory in logical form. *Annals of Pure and Applied Logic*, 51:1–77.

Abramsky, Samson (1991b). A domain equation for bisimulation. *Information and Computation*, 92(2):161–218.

Barr, M. and Wells, C. (1984). *Toposes, Triples and Theories*. Springer-Verlag. Reissued as Barr and Wells, 2005.

Barr, M. and Wells, C. (2005). *Toposes, Triples and Theories*. Number 12 in Reprints in Theory and Applications of Categories. Theory and Applications of Categories, Mount Allison University. Originally published as Barr and Wells, 1984.

Carboni, A., Pedicchio, M.C., and Rosolini, G., editors (1991). *Category Theory—Proceedings, Como 1990*, number 1488 in Lecture Notes in Mathematics. Springer-Verlag.

Coquand, T., Sambin, G., Smith, J., and Valentini, S. (2003). Inductively generated formal topologies. *Annals of Pure and Applied Logic*, 124:71–106.

Fourman, M.P. and Grayson, R.J. (1982). Formal spaces. In Troelstra and van Dalen, editors, *The L.E.J. Brouwer Centenary Symposium*, pages 107–122. North Holland.

Fourman, M.P. and Hyland, J.M.E. (1979). Sheaf models for analysis. In Fourman et al., 1979, pages 280–301.

Fourman, M.P., Mulvey, C.J., and Scott, D., editors (1979). *Applications of Sheaves*, number 753 in Lecture Notes in Mathematics. Springer-Verlag.

Ghilardi, Silvio (1989). Presheaf semantics and independence results for some nonclassical first-order logics. *Arch. Math. Logic*, 29(2):125–136.

Ghilardi, Silvio (1991). Incompleteness results in Kripke semantics. *J. Symbolic Logic*, 56(2):517–538.

Ghilardi, Silvio (1992). Quantified extensions of canonical propositional intermediate logics. *Studia Logica*, 51(2):195–214.

Gierz, G., Hofmann, K.H., Keimel, K., Lawson, J.D., Mislove, M., and Scott, D.S. (1980). *A Compendium of Continuous Lattices*. Springer-Verlag.

Goldblatt, Robert (1979). *Topoi: The Categorial Analysis of Logic*, volume 98 of *Studies in Logic and the Foundations of Mathematics*. North-Holland.

Goldblatt, Robert (1981). Grothendieck topology as geometric modality. *Z. Math. Logik Grundlag. Math.*, 27(6):495–529.

Isoda, Eiko (1997). Kripke bundle semantics and C-set semantics. *Studia Logica*, 58(3):395–401.

Johnstone, P.T. (1982). *Stone Spaces*. Number 3 in Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Johnstone, P.T. (1985). Vietoris locales and localic semi-lattices. In Hoffmann, R.-E., editor, *Continuous Lattices and their Applications*, number 101 in Pure and Applied Mathematics, pages 155–18. Marcel Dekker.

Johnstone, P.T. (2002a). *Sketches of an Elephant: A Topos Theory Compendium, vol. 1*. Number 44 in Oxford Logic Guides. Oxford University Press.

Johnstone, P.T. (2002b). *Sketches of an Elephant: A Topos Theory Compendium, vol. 2*. Number 44 in Oxford Logic Guides. Oxford University Press.

Johnstone, P.T. and Vickers, S.J. (1991). Preframe presentations present. In Carboni et al., 1991, pages 193–212.

Joyal, A. and Tierney, M. (1984). An extension of the Galois theory of Grothendieck. *Memoirs of the American Mathematical Society*, 309.

Lambek, J. and Scott, P.J. (1986). *Introduction to Higher-Order Categorical Logic*. Cambridge University Press.

Lavendhomme, R., Lucas, Th., and Reyes, G. (1989). Formal systems for topos-theoretic modalities. *Bull. Soc. Math. Belg. Sér. A*, 41(2):333–372.

Mac Lane, S. (1971). *Categories for the Working Mathematician*. Springer-Verlag.

Mac Lane, S. and Moerdijk, I. (1992). *Sheaves in Geometry and Logic*. Springer-Verlag.

Maietti, Maria Emilia (2003). Joyal's arithmetic universes via type theory. In Blute, Rick and Selinger, Peter, editors, *Category Theory and Computer Science (CTCS '02)*, number 69 in Electronic Notes in Theoretical Computer Science. Elsevier. doi:10.1016/S1571-0661(04)80569-3.

Makkai, M. and Reyes, G. (1995). Completeness results for intuitionistic and modal logic in a categorical setting. *Ann. Pure Appl. Logic*, 72(1):25–101.

McKinsey, J.C.C. and Tarski, A. (1944). The algebra of topology. *Ann. of Math.*, 45(2):141–191.

Nagaoka, K. and Isoda, E. (1997). Incompleteness results in Kripke bundle semantics. *Math. Logic Quart.*, 43(4):485–498.

Palmgren, Erik and Vickers, Steven (2007). Partial Horn logic and cartesian categories. *Ann. Pure. Appl. Logic*, 145(3):314–353.

Plotkin, G.D. (1981). Postgraduate lecture notes in advanced domain theory. Technical report, Dept of Computing Science, University of Edinburgh.

Reyes, G. and Zolfaghari, H. (1991). Topos-theoretic approaches to modality. In Carboni et al., 1991, pages 359–378.

Reyes, G. and Zolfaghari, H. (1996). Bi-Heyting algebras, toposes and modalities. *J. Philos. Logic*, 25(1):25–43.

Reyes, Gonzalo (1991). A topos-theoretic approach to reference and modality. *Notre Dame J. Formal Logic*, 32(3):359–391.

Robinson, E. (1986). Power-domains, modalities and the Vietoris monad. Technical Report 98, Computer Laboratory, University of Cambridge.

Sambin, G. (1987). Intuitionistic formal spaces—a first communication. In Skordev, Dimiter G., editor, *Mathematical Logic and its Applications*, pages 187–204. Plenum.

Shehtman, Valentin B. and Skvortsov, D.P. (1990). Semantics of non-classical first order predicate logics. In Petkov, P., editor, *Mathematical Logic*, pages 105–116. Plenum Press, New York.

Shirasu, Hiroyuki (1998). Duality in superintuitionistic and modal predicate logics. In Kracht, Marcus, de Rijke, Maarten, Wansing, Heinrich, and Zakharyaschev, Michael, editors, *Advances in Modal Logic, Volume 1 (Berlin, 1996)*, number 87 in CSLI Lecture Notes, pages 223–236. CSLI Publications, Stanford.

Skvortsov, D. and Shehtman, V. (1993). Maximal Kripke-type semantics for modal and superintuitionistic predicate logics. *Ann. Pure Appl. Logic*, 63(1): 69–101.

Skvortsov, Dmitrij (1996). On finite intersections of intermediate predicate logics. In Ursini, A. and P., Agliano, editors, *Logic and Algebra (Pontignano 1994)*, number 180 in Lect. Notes Pure Appl. Math., pages 667–688. Marcel Dekker.

Skvortsov, Dmitrij (2003). An incompleteness result for predicate extensions of intermediate propositional logics. In Balbiani, Philippe, Suzuki, Nobu-Yuki,

Wolter, Frank, and Zakharyaschev, Michael, editors, *Advances in Modal Logic, Volume 4*, pages 461–474. King's College Publications, London.

Smyth, M. (1978). Power domains. *Journal of Computer and System Sciences*, 16:23–36.

Suzuki, Nobu-Yuki (1990). Kripke bundles for intermediate predicate logics and Kripke frames for intuitionistic modal logics. *Studia Logica*, 49(3): 289–306.

Suzuki, Nobu-Yuki (1993). Some results on the Kripke sheaf semantics for superintuitionistic predicate logics. *Studia Logica*, 52(1):73–94.

Tarski, Alfred (1938). Der Aussagenkalcül und die Topologie. *Fund. Math.*, 31:103–134.

Vickers, S.J. (1995). Locales are not pointless. In Hankin, C.L., Mackie, I.C., and Nagarajan, R., editors, *Theory and Formal Methods of Computing 1994*, pages 199–216, Imperial College Press. London.

Vickers, S.J. and Townsend, C.F. (2004). A universal characterization of the double powerlocale. *Theoretical Computer Science*, 316:297–321.

Vickers, Steven (1989). *Topology via Logic*. Cambridge University Press.

Vickers, Steven (1997). Constructive points of powerlocales. *Math. Proc. Cam. Phil. Soc.*, 122:207–222.

Vickers, Steven (1999). Topical categories of domains. *Mathematical Structures in Computer Science*, 9:569–616.

Vickers, Steven (2001). Strongly algebraic = SFP (topically). *Mathematical Structures in Computer Science*, 11:717–742.

Vickers, Steven (2004). The double powerlocale and exponentiation: A case study in geometric reasoning. *Theory and Applications of Categories*, 12: 372–422.

Vickers, Steven (2005). Some constructive roads to Tychonoff. In Crosilla, Laura and Schuster, Peter, editors, *From Sets and Types to Topology and Analysis: Towards Practicable Foundations for Constructive Mathematics*, number 48 in Oxford Logic Guides, pages 223–238. Oxford University Press.

Vickers, Steven (2006). Compactness in locales and formal topology. *Annals of Pure and Applied Logic*, 137:413–438.

Viglas, K. (2004). *Topos Aspects of the Extended Priestley Duality*. PhD thesis, Department of Computing, Imperial College, London.

Wraith, G.C. (1979). Generic Galois theory of local rings. In Fourman et al., 1979, pages 739–767.

Chapter 9

SPATIAL LOGIC + TEMPORAL LOGIC = ?

Roman Kontchakov
Birkbeck College, London

Agi Kurucz
King's College, London

Frank Wolter
University of Liverpool

Michael Zakharyaschev
Birkbeck College, London

Second Reader

Philippe Balbiani
Institut de Recherche en Informatique de Toulouse

1. Introduction

As follows from the title of this chapter, our primary aim is to analyse possible solutions to the equation

$$\boxed{\textit{Spatial logic} \ + \ \textit{Temporal logic} \ = \ x} \tag{9.1}$$

where the items on the left-hand side are some standard *spatial* and *temporal logics*, and $+$ is some 'operator' combining these two logics into a single one. The question we are concerned with is how the computational complexity and the expressive power of the component logics are related to the complexity and expressiveness of the resulting *spatio-temporal logic* x under *various combination operators* $+$.

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 497–564.
© 2007 *Springer.*

To convey the flavour of the problems we are facing when attempting to answer this question, let us consider two standard spatial and temporal logics and try to combine them.

Recall from Ch. 5 of this Handbook that one of the basic and natural logics for reasoning about space is the 'modal' logic $\mathcal{S}4_u$ equipped with the Boolean operators over subsets of a topological space and the 'modal' operators **I** and **C** interpreted as the topological interior and closure, respectively. In this language we can say, for example, that two spatial objects X and Y are externally connected, $\mathsf{EC}(X, Y)$ in symbols, in the sense that X and Y share some points but none of them belongs to the interior of X or Y. This can be expressed, e.g., by means of the following constraints:

$$X \cap Y \neq \emptyset \quad \text{and} \quad \mathbf{I}X \cap \mathbf{I}Y = \emptyset.$$

Reasoning in $\mathcal{S}4_u$ is perfectly well understood; it is known to be PSPACE-complete, and various reasonably effective reasoning systems are available.

For the temporal component we take the standard linear temporal logic $\mathcal{LTL}$ which extends propositional logic with the temporal operators $\bigcirc$ ('tomorrow'), $\Diamond_F$ (eventually), and $\Box_F$ (always in the future). $\mathcal{LTL}$ is interpreted over the flow of time consisting of the natural numbers $(\mathbb{N}, <)$. For example, the following formula says that a day is Saturday if, and only if, the next day is Sunday:

$$\Box_F(\mathit{Saturday} \leftrightarrow \bigcirc \mathit{Sunday}).$$

Reasoning in $\mathcal{LTL}$ is also thoroughly investigated; it is PSPACE-complete as well, and a number of temporal reasoning systems have been implemented.

Now our aim is to construct a combination of $\mathcal{S}4_u$ and $\mathcal{LTL}$ where we could express, for example, that today spatial objects X and Y are not externally connected, but tomorrow they are:

$$\neg\mathsf{EC}(X, Y) \ \wedge \ \bigcirc\mathsf{EC}(X, Y),$$

or that the spatial object X today is externally connected with the space $\bigcirc X$ it will be occupying tomorrow:

$$\mathsf{EC}(X, \bigcirc X),$$

or that, starting from some future moment, X will never change its position:

$$\Diamond_F \Box_F (X = \bigcirc X).$$

Having efficient spatial and temporal reasoners S and T at our disposal (for $\mathcal{S}4_u$ and $\mathcal{LTL}$, respectively), the quickest way of constructing a combined spatio-temporal reasoning system is to organise their joint work in a modular way: first, say, S treats the input constraints regarding formulas that start with temporal

operators as atomic, then T deals with them regarding formulas with spatial operators as atomic, etc. Clearly, the resulting system works in PSPACE. But unfortunately, such a reasoner does not take into account any interaction between the spatial and temporal operators: the problem is that a spatio-temporal formula is recognised as valid by this reasoner only if it is valid in *arbitrary fusions of topological models with* (*possibly many*) *isomorphic copies of the flow of time* $(\mathbb{N}, <)$. In such models, spatial objects are not moving in *the same space* over *the same flow of time* because the topological space at moment n may have absolutely nothing to do with the space at moment $n + 1$, or, dually, every point of space has its own history. In particular, one could expect the constraint

$$\bigcirc\mathsf{EC}(X, Y) \leftrightarrow \mathsf{EC}(\bigcirc X, \bigcirc Y)$$

to be a valid principle of spatial-temporal logic—yet, our reasoner would not confirm this: it would claim that the negation of this formula is satisfiable.

Of course, from a purely semantical perspective, this problem can easily be overcome by restricting the class of *intended models* to those where the same topological space is kept along the whole time line. In other words, we can assume that the underlying topological space does not change in time; what changes is the position, shape, size, etc. of spatial objects. Mathematically this means that the intended spatio-temporal models for combinations of $\mathcal{S}4_u$ and $\mathcal{LTL}$ are the Cartesian products of topological spaces and $(\mathbb{N}, <)$.

Such models provide a natural interpretation for the formulas considered above, with the last one being valid in all of them. But on the other hand, in order to deal with them we need a new, perhaps more sophisticated reasoning system. Is it, at least in principle, possible to design an effective complete and sound system of this kind?

A moment's reflection about the possible computational behaviour of such a system brings to memory another model, which logicians and computer scientists know all too well. We mean Turing's model of computation. The tape of a Turing machine can be regarded as a somewhat simplified model of space where a 'spatial object' is the collection of cells containing a certain symbol from the alphabet. Putting the problem in this perspective, one can immediately start suspecting that perhaps even a modestly expressive spatio-temporal language could be able to describe the change of spatial objects over time which corresponds to the computation of a Turing machine. And if this is indeed the case then, using the operator $\Diamond_F$ for 'eventually' it appears almost trivial to state that the Turing machine eventually reaches the halting state on a given input, which would mean that reasoning in the hybrid language cannot be decidable (or, even worse, that the set of valid spatio-temporal formulas is not recursively enumerable).

Now, obviously, the topological language $\mathcal{S}4_u$ and many other languages to be considered in this chapter are not designed to represent knowledge about the

tape of a Turing machine (to begin with, there is no obvious topology on such a tape). Some much smarter 'encoding techniques' may be needed to prove that combinations of $\mathcal{S}4_u$ and $\mathcal{LTL}$ (and similar logics) are undecidable. Yet, the *first major result* of this chapter shows that the intuition behind the simulation of Turing machines discussed above is correct: naïve and straightforward combinations of spatial and temporal logics (interpreted in Cartesian products of time and space) almost invariably lead to undecidable hybrids.

The *second major result*, however, is that by closely inspecting the expressive means required to simulate Turing machines one can still find hierarchies of useful and expressive hybrids of $\mathcal{S}4_u$ and $\mathcal{LTL}$, their fragments, and some related logics which are decidable and of reasonably low complexity.

The structure of this chapter is as follows. In the next section, we discuss in more detail, but still on a rather abstract level, our main paradigm of 'snapshot spatio-temporal models' and most important reasoning problems relevant to these models.

Then, in Sec. 3 and 4, we discuss in detail the ingredients of the spatio-temporal logics to be constructed and investigated in this chapter. We consider two families of spatial logics. The first one is comprised of formalisms designed for reasoning about topological relations among spatial objects and ranging from $\mathcal{RCC}$-8 to $\mathcal{S}4_u$, possibly with component counting. A remarkable feature of these logics is their 'computational robustness' in the sense that the complexity of reasoning gradually increases from NP for $\mathcal{RCC}$-8 to PSPACE and NEXPTIME for $\mathcal{S}4_u$ without and with component counting, respectively. Moreover, each complexity 'jump' in this hierarchy is clearly connected to the corresponding increase in the logic's expressiveness. Our second family of spatial logics consists of formalisms that are capable of reasoning about distances in metric spaces. Some of these logics will contain $\mathcal{S}4_u$ and, therefore, combine topological reasoning with reasoning about distances. These logics are also computationally robust, with the typical complexity being EXPTIME.

The introduction to temporal logic systems in Sec. 4 is much shorter, as we only consider two approaches to logic modelling of time: time as a linear discrete sequence of time points or snapshots, and time as a tree-like structure of such snapshots representing some aspects of non-determinism. Other flows of time, say, continuous time, are not discussed, but pointers to the literature are provided.

Having introduced the logical systems for space and time, in Sec. 5 we discuss general combination principles—requirements and constraints for the operator $+$ in (9.1)—which will guide us when designing combined spatio-temporal systems. Then, in Sec. 6 and 7, we use these principles to construct spatio-temporal logics out of the components introduced in Sec. 3 and 4. As before, the emphasis of this investigation is on the trade-off between the expressive power and the complexity of reasoning. We shall discover, in particular,

that unlike the 'robust' component logics, their spatio-temporal hybrids turn out to be much more sensitive to seemingly minor changes in expressiveness.

In Sec. 8, we consider a somewhat different paradigm of spatio-temporal models and languages for reasoning about them: here we formalise spatio-temporal reasoning within the framework of dynamical systems based on topological and metric spaces with continuous and isometric functions, respectively. As logics for dynamical systems are discussed in detail elsewhere in this Handbook (see Ch. 10), we concentrate here on the connection between the spatio-temporal systems introduced before and the dynamical systems perspective. It will turn out that in some cases the connections between the two approaches are so strong that results can be mutually imported from one area to the other.

Finally, in Sec. 9, we briefly discuss the relation between spatio-temporal logics and other temporalised formalisms, for example first-order temporal logics and temporal epistemic logics.

The reader who considers computational complexity less important and is interested in logic modelling of (relativistic) space-time using classical *first-order* logic is referred to Ch. 11.

2. Static and changing spatial models

The intended models of standard spatial logics are usually based on 'mathematical spaces' such as (variations of) topological or metric spaces and their relational or algebraic representations or abstractions. We will consider many examples of such models and spaces in Sec. 3.1; more can be found elsewhere in this Handbook. Meanwhile, in order to discuss basic principles of introducing a temporal dimension into otherwise static spatial models, we neglect the concrete structure of these 'mathematical spaces' and concentrate on the generic properties of the models.

To represent spatial entities in models we require a countably infinite supply of *spatial variables* (that is, unary predicates) $p_0, p_1, \dots$. Thus, a generic spatial model can be thought of as a structure of the form

$$\mathfrak{M} = (\mathfrak{S}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots), \tag{9.2}$$

where $\mathfrak{S}$ is the underlying 'mathematical space' (say, a metric or topological space, or a structured collection of polygons on the Euclidean plane) and the $p_i^{\mathfrak{M}}$ are interpretations of the spatial variables as subsets of the domain of $\mathfrak{S}$.

Depending on the underlying spatial ontology, one can distinguish between two types of models:

- *point-based* models, where spatial objects are (explicitly or implicitly) thought of as consisting of sets of points, and
- models with *extended spatial entities* as basic elements (say, regions or intervals) together with certain relations between them.

Point-based spatial models. In a point-based model of the form (9.2), the underlying 'mathematical space' $\mathfrak{S}$ is a collection of points equipped with 'point-wise defined' operators (like a metric or topological space). Interpretations $p_i^{\mathfrak{M}}$ of spatial variables p_i (that is, subsets of the domain of $\mathfrak{S}$) represent spatial objects. Thus, a spatial object is identified with the set of points it occupies. By imposing various constraints on these interpretations—say, by allowing only polygons, circles or regular closed connected sets—we can reflect the desired requirements on the form of spatial objects.

Region-based spatial models. In a region-based model of the form (9.2), spatial objects are represented as (unstructured) *elements* of the underlying 'space' $\mathfrak{S}$. We may consider as the domain of $\mathfrak{S}$, for instance, the collection of polygons on the Euclidean plane and completely forget about the plane itself. The 'structure' of spatial objects is reflected then by certain relations among them (say, polygon x has a common edge with polygon y) which should be specified in $\mathfrak{S}$ (for details and further references see Ch. 2, 3 and 7 of this Handbook). Spatial variables are again interpreted as sets of elements of the domain of $\mathfrak{S}$, for instance as a set of polygons approximating the map of the U.K. (including the Isle of Wight, the Hebrides, and other islands), or the singleton set containing (the polygonal approximation of) the Isle of Man.

In this chapter we only consider point-based spatial models, although some results and constructions can be generalised to region-based ones.

Snapshot spatio-temporal models. The intended models of temporal logics are supposed to represent the *change of states*—which, in our case, should be spatial models (9.2)—over time, under actions, etc. In most cases it makes sense to assume that space always remains the same. Moreover, one can usually simulate expanding, shrinking or varying space in some 'sufficiently large' constant space (e.g., Gabbay et al., 2003). The motion of spatial objects can therefore be modelled by changing the interpretations $p_i^{\mathfrak{M}}$ of spatial objects from one state to another. (A different approach to modelling motion was taken by Muller (1998), who considered a moving object as a single spatio-temporal entity.)

There are many different time paradigms developed in philosophy, mathematics, physics, computer science and other disciplines: linear and branching, discrete and dense, point-based and interval-based, etc. (e.g., Gabbay et al., 1994; Gabbay et al., 2000; Fisher et al., 2005). In this chapter we mainly focus on the flow of time that can be represented by the natural numbers $(\mathbb{N}, <)$, where $<$ is the temporal precedence relation between time points. In this case our generic *snapshot spatio-temporal model* is simply an infinite sequence

$$\mathfrak{M}_0 = (\mathfrak{S}, p_0^{\mathfrak{M}_0}, p_1^{\mathfrak{M}_0}, \dots), \quad \mathfrak{M}_1 = (\mathfrak{S}, p_0^{\mathfrak{M}_1}, p_1^{\mathfrak{M}_1}, \dots), \quad \dots \tag{9.3}$$

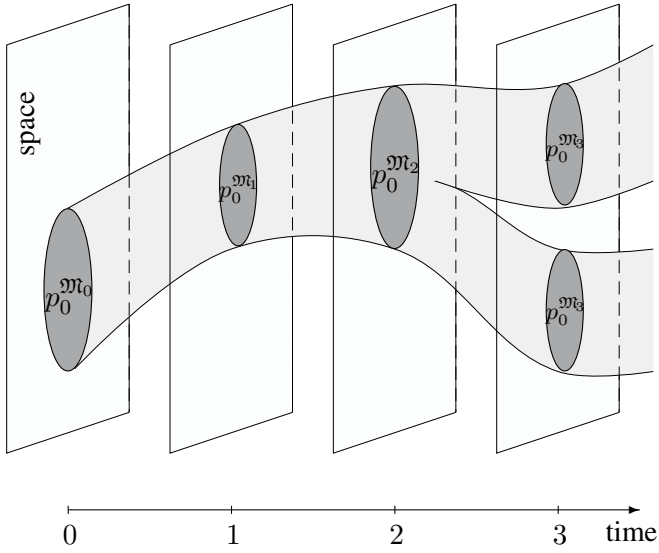


Figure 9.1. Linear snapshot model with a moving spatial object.

of spatial models of the form (9.2) with the same space $\mathfrak{S}$; see Fig. 9.1. In Sec. 4.2 and 6.2 we will briefly consider temporal and spatio-temporal models with branching (tree-like discrete) time that can capture some aspects of non-determinism. In either of these time paradigms the points of time can be taken as primitive temporal entities, assumed to be generated by state transition systems (a standard computer science approach), or by dynamical systems (a usual way in mathematics).

As an illustration let us consider the following example.

Spatial transition systems. Our main example running throughout the chapter is a spatial transition system which describes the changing geography of the Earth as we see it every day in BBC's weather forecasts, say, in Ten O'Clock News. Every day the state of the map is represented by a spatial model

$$\mathfrak{M} = (\mathfrak{E}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots),$$

where $\mathfrak{E}$ is a suitable mathematical model of the Earth surface and each $p_i^{\mathfrak{M}}$ is the space occupied by the geographical object modelled by p_i (either static as a town, a county or dynamic as a night frost or rainfall area, etc.) on that day. Starting from a certain day in the past, we can trace then the day-after-day changes that have happened till the present moment. Depending on our philosophical, religious, etc. views we can regard the future to be deterministic

or not. In particular, we can imagine that today's state may evolve in many different ways.

In computer science, such scenarios are often described in terms of state transition systems—*spatial transition systems* in our context—which, in general, are tuples of the form

$$(S, \rightarrowtail, \mu, \mathfrak{S}, s_0), \tag{9.4}$$

where S is a nonempty set of *states*, $\rightarrowtail$ is a binary *transition* relation on S without dead-ends (states without outgoing $\rightarrowtail$), μ is a function associating with each state $s \in S$ a spatial model $\mu(s)$ of the form (9.2) based on the same space $\mathfrak{S}$, and s_0 is the *initial state*. Possible *evolutions* (or *transformations*) of this initial state are sequences

$$s_0 \rightarrowtail s_1 \rightarrowtail s_2 \rightarrowtail \dots, \tag{9.5}$$

where $s_i \in S$ for all $i \in \mathbb{N}$. Each of these evolutions obviously generates a linear snapshot model

$$\mu(s_0),\ \mu(s_1),\ \mu(s_2),\ \dots$$

of the form (9.3). In the deterministic case (as in the second example below) we have a single evolution. In general, however, the transition relation $\rightarrowtail$ of a spatial transition system can represent non-deterministic rules. Then it generates a discrete tree of evolutions (9.5).

What precisely can be told about these models depends of course on the concrete spatial and temporal logics we use. Here we give a few examples of English statements about our 'geographical transition system' that will be represented as spatio-temporal formulas in Sec. 5 and 6.1.

(A) If two clouds are disconnected now, then at the next moment they either remain disconnected or become externally connected.

(B) Kaliningrad is disconnected from the EU until the moment when Poland becomes a tangential proper part of the EU, after which Kaliningrad and the EU are externally connected forever.

(C) The current position of a hurricane overlaps its position in an hour.

(D) If tomorrow object X is at the place where object Y is today, then Y will have to move by tomorrow.

(E) The space occupied by Europe never changes.

(F) In two years the EU will be extended with Romania and Bulgaria.

(G) It will be raining over every part of England ever and ever again.

(H) If the Earth consists of water and land, and the space occupied by water expands, then the space occupied by land shrinks.

(I) Two deserts that expand by at least a mile in all directions every year must eventually intersect.

Reasoning tasks. We have not introduced yet any formal languages capable of talking and reasoning about spatio-temporal models—they depend on the concrete spatial and temporal logics we combine as well as the combination principles to be discussed later on in Sec. 5. Nonetheless, it does make sense to consider on this abstract level the main *reasoning problems* one might be interested in for some fictional language $\mathcal{L}$.

The most general and important problem we are going to consider is

- *satisfiability of spatio-temporal constraints*.

Suppose that we have formulated a finite set Γ of $\mathcal{L}$-formulas representing constraints on possible spatio-temporal scenarios. Then we are facing the following questions. Is this set Γ satisfiable (or consistent)? In other words, does there exist a spatio-temporal model realising these constraints? And if so, how such a model may look like? For example, can it be given by a finite transition system? Can it be based on a finite space?

Of particular interest to us will be algorithmic properties of the satisfiability problem. Is this problem *decidable*? That is, does there exist an algorithm which is capable of deciding, given an arbitrary finite set Γ of constraints, whether Γ is satisfiable? Are finite sets of satisfiable constraints *recursively enumerable*? What is the computational (worst-case) complexity of the satisfiability problem?

Note that the *deduction* (or *entailment*) *problem* 'given a finite set Γ of constraints and an $\mathcal{L}$-formula φ, decide whether φ holds in all spatio-temporal models where Γ holds?' is usually reducible to the satisfiability problem.

The satisfiability problem can be restricted to certain classes of $\mathcal{L}$-formulas and constraints. Here is a typical example. We describe the behaviour of spatial transition systems by imposing some *local* constraints Γ which specify possible initial states and transitions from each given state to the next ones. This is done in some sublanguage $\mathcal{L}_{loc}$ of $\mathcal{L}$. We can also specify (by means of $\mathcal{L}_{loc}$-formulas) states with some desirable property φ or some 'bad' property ψ. And then we are interested in the algorithmic properties of

- the *reachability problem* relative to $\mathcal{L}_{loc}$: 'is it the case that every model where constraints Γ hold contains a state satisfying φ?' or
- the *safety problem* relative to $\mathcal{L}_{loc}$: 'is it the case that no model where constraints Γ hold contains a state satisfying ψ?'

In the extreme case, when $\mathcal{L}_{loc}$ is expressive enough to describe (up to isomorphism) any particular spatial transition system, checking for reachability,

safety or some other properties are instances of classical *model checking* problems (see, e.g., Clarke et al., 2000, and references therein).

3. Spatial logics

In this section we introduce the 'mathematical spaces' and spatial logics capable of talking and reasoning about these spaces that will serve as the spatial components of our spatio-temporal formalisms.

We consider spatial logics of two types: (i) those that can represent and reason about topological relations among spatial objects, and (ii) those that can additionally take into account distances between objects. The former are interpreted over *topological spaces* and the latter over *metric* (or more generally, *distance*) *spaces*. The choice of these logics is motivated by the following reasons. First, topological and metric spaces belong to the most important and well-understood structures representing space. Reasoning about topological relations between regions such as 'X is externally connected to Y' or 'X is tangential proper part of Y' has proved to be one of the most successful approaches to *qualitative* spatial knowledge representation and reasoning (KR&R) in artificial intelligence; see, e.g., (Cohn and Hazarika, 2001) and references therein. Extensions of 'topologics' with distance operators like $\exists^{\leq a} X$ giving the a-neighbourhood of X or $X \leftleftarrows Y$ giving the set of points that are closer to X than to Y are becoming another interesting research stream (Kutz et al., 2003; Wolter and Zakharyaschev, 2003; Wolter and Zakharyaschev, 2005a) that is especially close to the authors' hearts. Other important aspects of space such as, e.g., orientation have been considered as well; however, no combinations with temporal logics have been constructed so far. We believe that the approach to combining spatial logics of topological and metric spaces with temporal ones to be presented later on in this chapter can be extended to other spatial formalisms as well.

3.1 Metric and topological spaces

Metric spaces. A *metric space* is a pair (Δ, d), where Δ is a nonempty set (of points) and d is a function from $\Delta \times \Delta$ into the set $\mathbb{R}^{\geq 0}$ (of non-negative real numbers) satisfying the following axioms

(9.6) *identity of indiscernibles*: $d(x, y) = 0 \quad \text{iff} \quad x = y,$

(9.7) *symmetry*: $d(x, y) = d(y, x),$

(9.8) *triangle inequality*: $d(x, z) \leq d(x, y) + d(y, z),$

for all $x, y, z \in \Delta$. The value $d(x, y)$ is called the *distance* between points x and y. Given a metric space (Δ, d), a point $x \in \Delta$ and a nonempty $Y \subseteq \Delta$,

define the *distance* $d(x, Y)$ between x and Y by taking

$$d(x, Y) = \inf\{d(x, y) \mid y \in Y\}.$$

As usual, $d(y, \emptyset) = \infty$. The distance $d(X, Y)$ between two nonempty sets X and Y is

$$d(X, Y) = \inf\{d(x, y) \mid x \in X,\ y \in Y\}.$$

Distance spaces. Although acceptable in many cases, the defined concept of metric space is not universally applicable to all interesting measures of distance between points, especially those used in everyday life. Consider, for instance, the following two examples:

(i) If $d(x, y)$ is the flight-time from x to y then, as we know it too well, d is not necessarily symmetric, even approximately (just take a plane from London to Tokyo and back).

(ii) Often we do not measure distances by means of real numbers but rather using more fuzzy notions such as 'short,' 'medium' and 'long.' To represent these measures we can, of course, take functions d from $\Delta \times \Delta$ into the subset $\{1, 2, 3\}$ of $\mathbb{R}^{\geq 0}$ and define $short := 1$, $medium := 2$, and $long := 3$. So we can still regard these distances as real numbers. However, for measures of this type the triangle inequality (9.8) does not make sense (short plus short can still be short, but it can also be medium or long).

Spaces (Δ, d) satisfying only the axiom (9.6) are called *distance spaces*.

Topological spaces. A *topological space* is a pair $(U, \mathbb{I})$ in which U is a nonempty set, the *universe* of the space, and $\mathbb{I}$ is the *interior operator* on U satisfying the *Kuratowski axioms*: for all $X, Y \subseteq U$,

$$\mathbb{I}(X \cap Y) = \mathbb{I}X \cap \mathbb{I}Y, \qquad \mathbb{I}X \subseteq \mathbb{I}\mathbb{I}X, \qquad \mathbb{I}X \subseteq X \quad \text{and} \quad \mathbb{I}U = U.$$

The operator dual to $\mathbb{I}$ is called the *closure operator* and denoted by $\mathbb{C}$: for every $X \subseteq U$, we have $\mathbb{C}X = U - \mathbb{I}(U - X)$. Thus, $\mathbb{I}X$ is the *interior* of a set X, while $\mathbb{C}X$ is its *closure*. X is called *open* if $X = \mathbb{I}X$ and *closed* if $X = \mathbb{C}X$. The complement of an open set is closed and vice versa. The *boundary* of a set $X \subseteq U$ is defined as $\mathbb{C}X - \mathbb{I}X$. Note that X and $U - X$ have the same boundary.

Topological spaces are often (equivalently) defined as pairs $(U, \mathcal{O})$, where $\mathcal{O}$ is a family of (open) subsets of U such that $\mathcal{O}$ is closed under arbitrary unions and finite intersections.

Metric spaces and topology. Each metric space (Δ, d) gives rise to the *interior operator* $\mathbb{I}_d$ on Δ: for all $X \subseteq \Delta$,

$$\mathbb{I}_d X = \{x \in X \mid \exists \varepsilon > 0\ \forall y\ (d(x, y) < \varepsilon \rightarrow y \in X)\}.$$

The pair $(\Delta, \mathbb{I}_d)$ is called the *topological space induced by* the metric space (Δ, d). The dual *closure operator* $\mathbb{C}_d$ in this space can be defined by the equality

$$\mathbb{C}_d X = \{x \in \Delta \mid \forall \varepsilon > 0\ \exists y \in X\ d(x, y) < \varepsilon\}.$$

We briefly remind the reader of a few standard examples of metric and topological spaces that will be used in what follows.

Euclidean spaces. The *one-dimensional Euclidean space* is the set of real numbers $\mathbb{R}$ equipped with the following metric on it

$$d_1(x, y) = |x - y|.$$

Let $X \subseteq \mathbb{R}$. A point $x \in \mathbb{R}$ is said to be *interior* in X if there is some $\varepsilon > 0$ such that the whole open interval $(x - \varepsilon, x + \varepsilon)$ belongs to X. The interior $\mathbb{I}X$ of X is defined then as the set of all interior points in X. It is not hard to check that $(\mathbb{R}, \mathbb{I})$ is the topological space induced by the Euclidean metric d_1. Open sets in $(\mathbb{R}, \mathbb{I})$ are (possibly infinite) unions of open intervals (a, b), where $a \leq b$. The closure of (a, b), for $a < b$, is the closed interval $[a, b]$, with the end points a and b being its boundary.

In the same manner one can define *n-dimensional Euclidean spaces* based on the universes $\mathbb{R}^n$ with the metric

$$d_n(x, y) = \sqrt{\textstyle\sum_{i=1}^{n} (x_i - y_i)^2},$$

where x and y are n-dimensional vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively (in the definition of interior points x one should take n-dimensional ε-neighbourhoods of x).

Metric spaces on graphs. Another well known example is *metric spaces on graphs*: the distance between two nodes of a graph is defined as the length of the shortest path between them.

Aleksandrov spaces. A topological space is called an *Aleksandrov space* (Alexandroff, 1937) if arbitrary (not only finite) intersections of open sets are open. Aleksandrov spaces are closely related to *quasi-ordered sets*, that is, pairs $\mathfrak{G} = (V, R)$, where V is a nonempty set and R a transitive and reflexive relation on V. Every such quasi-order $\mathfrak{G}$ induces the interior operator $\mathbb{I}_{\mathfrak{G}}$ on V: for $X \subseteq V$,

$$\mathbb{I}_{\mathfrak{G}} X = \{x \in X \mid \forall y \in V\ (xRy \to y \in X)\}.$$

In other words, the open sets of the topological space $\mathfrak{T}_{\mathfrak{G}} = (V, \mathbb{I}_{\mathfrak{G}})$ are the *upward closed* (or *R-closed*) subsets of V. The *minimal neighbourhood* of a point x in $\mathfrak{T}_{\mathfrak{G}}$ (that is, the minimal open set to contain x) consists of all those

points that are R-accessible from x. It is well-known (e.g., Bourbaki, 1966) that $\mathfrak{T}_{\mathfrak{G}}$ is an Aleksandrov space and, conversely, every Aleksandrov space is induced by a quasi-order.

For various generalisations of metric and topological spaces (like semi-metrics, closure spaces and digital topology) see Ch. 12.

3.2 Topo-logics

In this section, we introduce and discuss a number of logical formalisms which can represent and reason about topological relations among spatial objects interpreted over topological spaces. Our choice of logics was guided by two criteria: (i) they should be sufficiently expressive to represent interesting and useful topological knowledge as identified in the qualitative spatial reasoning community; (ii) on the other hand, reasoning with such logics should be decidable and, if possible, of low computational complexity. Another important constraint on logics in the framework described in Sec. 2 is that change in time is modelled by changing the extensions of *unary predicates* representing spatial objects.

The most developed and systemically studied spatial logics satisfying our criteria are fragments of a 'propositional' logic in which 'propositional variables' (= unary predicates) denote spatial objects, and topological relations among them are represented by means of the interior and closure operators, the universal and existential quantifiers over space, as well as the Booleans. This logic, originally introduced as a *modal* logic, is known as $\mathcal{S}4_u$. As we shall see below, it can be regarded as the logic of topological spaces providing a common roof to some other formalisms developed by the spatial community such as the $\mathcal{RCC}$-8 or 9-intersection region connection calculi (where topological relations between regions are regarded as primitive).

Our exposition basically follows Gabelaia et al., 2005a, where the reader can find more details, references and proofs. For historical references and motivation see Ch. 5 of this Handbook.

Modal logic of topological spaces. $\mathcal{S}4_u$ is the well known *propositional modal logic* $\mathcal{S}4$ extended with the universal modalities. The 'pedigree' of $\mathcal{S}4$ is quite unusual. It was introduced independently by Orlov (1928), Lewis (Lewis and Langford, 1932), and Gödel (1933), without any intention to reason about space. Orlov and Gödel understood it as a logic of 'provability' (in order to provide a classical interpretation for the intuitionistic logic of Brouwer and Heyting) and Lewis as a logic of necessity and possibility, that is, as a *modal logic*. That it can be regarded as the logic of topological spaces was discovered by Stone (1937), Tarski (1938), Tsao-Chen (1938) and McKinsey (1941).

In the spatial context it is useful to distinguish between spatial terms and spatial formulas of $\mathcal{S}4_u$ as explained below. *Spatial terms* are expressions of the form:

$$\tau \quad ::= \quad p_i \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \tau_1 \sqcup \tau_2 \quad | \quad \mathbf{I}\tau \quad | \quad \mathbf{C}\tau, \tag{9.9}$$

where

- the p_i are *spatial variables*,
- $\overline{}$, $\sqcap$ and $\sqcup$ are the standard Boolean operators (to be interpreted by the set-theoretic complement, intersection and union),
- **I** and **C** are the *interior* and *closure operators*, respectively (they correspond to the box and diamond of the modal logic $\mathcal{S}4$ but are denoted differently to emphasise their topological nature).

A *topological model* is a structure of the form

$$\mathfrak{M} = (\mathfrak{T}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots), \tag{9.10}$$

where $\mathfrak{T} = (U, \mathbb{I})$ is a topological space and $p_i^{\mathfrak{M}} \subseteq U$ for all i. The *extension* (or *interpretation*) $\tau^{\mathfrak{M}}$ of an arbitrary spatial term τ in $\mathfrak{M}$ is defined inductively by taking:

$$\overline{\tau}^{\mathfrak{M}} = U - \tau^{\mathfrak{M}}, \qquad (\tau_1 \sqcap \tau_2)^{\mathfrak{M}} = \tau_1^{\mathfrak{M}} \cap \tau_2^{\mathfrak{M}}, \qquad (\mathbf{I}\tau)^{\mathfrak{M}} = \mathbb{I}\tau^{\mathfrak{M}},$$
$$(\tau_1 \sqcup \tau_2)^{\mathfrak{M}} = \tau_1^{\mathfrak{M}} \cup \tau_2^{\mathfrak{M}} \qquad \text{and} \qquad (\mathbf{C}\tau)^{\mathfrak{M}} = \mathbb{C}\tau^{\mathfrak{M}}.$$

To be able to express how spatial terms τ_1 and τ_2 are related to each other we require (at least) the atomic formula $\tau_1 \sqsubseteq \tau_2$ with the obvious intended meaning: (the extension of) τ_1 is a subset of (the extension of) τ_2. By taking Boolean combinations of such atoms, we arrive at what will be called *spatial formulas* (or *$\mathcal{S}4_u$-formulas*):

$$\varphi \quad ::= \quad \tau_1 \sqsubseteq \tau_2 \quad | \quad \neg\varphi \quad | \quad \varphi_1 \wedge \varphi_2 \quad | \quad \varphi_1 \vee \varphi_2,$$

where the τ_i are spatial terms. Formally, the language of $\mathcal{S}4_u$ as defined above is weaker than the standard one, say, from Goranko and Passy, 1992. However, one can easily show that they have precisely the same expressive power: see, e.g., Hughes and Cresswell, 1996 or Aiello and van Benthem, 2002.

Spatial formulas can be either true or false in topological models. The *truth-relation* $\mathfrak{M} \models \varphi$—a spatial formula φ is true in a topological model $\mathfrak{M}$—is defined in the following way:

- $\mathfrak{M} \models \tau_1 \sqsubseteq \tau_2$ iff $\tau_1^{\mathfrak{M}} \subseteq \tau_2^{\mathfrak{M}}$,
- $\mathfrak{M} \models \neg\varphi$ iff $\mathfrak{M} \not\models \varphi$,

- $\mathfrak{M} \models \varphi_1 \wedge \varphi_2$ iff $\mathfrak{M} \models \varphi_1$ and $\mathfrak{M} \models \varphi_2$,
- $\mathfrak{M} \models \varphi_1 \vee \varphi_2$ iff $\mathfrak{M} \models \varphi_1$ or $\mathfrak{M} \models \varphi_2$.

Clearly, the traditional *universal modalities* $\forall$ and $\exists$ of $\mathcal{S}4_u$ are expressible in the above language: $\forall\tau$ can be regarded as an abbreviation for $(\top \sqsubseteq \tau)$ and $\exists\tau$ for $\neg(\tau \sqsubseteq \bot)$, where $\top$ and $\bot$ are constant terms denoting the whole space and the empty set, respectively. In what follows we will also freely use two other 'atomic' formulas $\tau_1 = \tau_2$ and $\tau_1 \neq \tau_2$ standing for $(\tau_1 \sqsubseteq \tau_2) \wedge (\tau_2 \sqsubseteq \tau_1)$ and $\neg(\tau_1 = \tau_2)$, respectively.

Say that a spatial formula φ is *satisfiable* (*in a class $\mathcal{K}$ of topological models*) if there is a topological model $\mathfrak{M}$ (from $\mathcal{K}$) such that $\mathfrak{M} \models \varphi$. A spatial formula φ is *satisfiable in a class of topological spaces* if there is a topological model $\mathfrak{M}$ based on a space from this class such that $\mathfrak{M} \models \varphi$.

This seemingly simple spatial language $\mathcal{S}4_u$ can express rather complex relations between sets in topological spaces. For example, the formula

$$(q \sqsubseteq p) \quad \wedge \quad (p \sqsubseteq \mathbf{C}q) \quad \wedge \quad (p \neq \bot) \quad \wedge \quad (\mathbf{I}q = \bot)$$

says that a set q is dense in a nonempty set p, but has no interior. As an example one can take q to be the rationals $\mathbb{Q}$ and p to be $\mathbb{R}$ in the Euclidean space $(\mathbb{R}, \mathbb{I})$.

In the following theorem we collected the most important facts about $\mathcal{S}4_u$; for proofs and discussions see, e.g., Nutt, 1999, Areces et al., 2000 and references therein.

THEOREM 9.1 (i) *A spatial formula is satisfiable iff it is satisfiable in an Aleksandrov space.*

(ii) *$\mathcal{S}4_u$ enjoys the exponential finite model property in the sense that every satisfiable spatial formula φ is satisfiable in a topological space whose size is at most exponential in the size of φ.*

(iii) *Satisfiability of spatial formulas in topological models is* PSPACE-*complete.*

The language of the modal logic $\mathcal{S}4$ mentioned above coincides with the language of $\mathcal{S}4_u$-terms. Say that a spatial term (= $\mathcal{S}4$-formula) is *satisfiable* if there is a topological model where the term is interpreted as a nonempty set. Although being of the same computational complexity as $\mathcal{S}4$ (which is also PSPACE-complete), the logic $\mathcal{S}4_u$ is more expressive. For example, spatial formulas can distinguish between arbitrary and connected topological spaces (we remind the reader that a topological space is *connected* if its universe cannot be represented as the union of two disjoint nonempty open sets). Consider the formula

$$(\mathbf{C}p \sqsubseteq p) \quad \wedge \quad (p \sqsubseteq \mathbf{I}p) \quad \wedge \quad (p \neq \bot) \quad \wedge \quad (p \neq \top) \tag{9.11}$$

Figure 9.2. Regular closure.

saying that (the extension of) p is both closed and open, nonempty and does not coincide with the whole space. It can only be satisfied in a model based on a disconnected topological space, while all satisfiable $\mathcal{S}4$-terms are satisfied in connected (e.g., Euclidean) spaces. For we have the following result (McKinsey and Tarski, 1944):

THEOREM 9.2 *An $\mathcal{S}4$-formula is satisfiable iff it is satisfiable in any of* (*and so in all*) $\mathbb{R}^n$, $n > 0$.

Another example illustrating the expressive power of $\mathcal{S}4_u$ is the formula

$$(p \neq \bot) \quad \wedge \quad (p \sqsubseteq \mathbf{C}\overline{p}) \quad \wedge \quad (\overline{p} \sqsubseteq \mathbf{C}p) \tag{9.12}$$

defining a nonempty set p such that both p and its complement $\overline{p}$ have empty interiors. In fact, the second and the third conjuncts say that both p and $\overline{p}$ consist of boundary points only.

Regions = regular closed sets. In qualitative spatial KR&R, it is quite often assumed that spatial terms can only be interpreted by regular closed (or open) sets of topological spaces (e.g., Davis, 1990; Asher and Vieu, 1995; Gotts, 1996). One of the reasons for imposing this restriction is to exclude from consideration such 'pathological' sets as in (9.12). Recall that a set X is *regular closed* if $X = \mathbb{C}\mathbb{I}X$, which clearly does not hold for any set satisfying (9.12). Another reason is to ensure that the space occupied by a physical body is homogeneous in the sense that it does not contain parts of 'different dimensionality.' For example, the one-dimensional curve in Fig. 9.2 disappears from the subset X of the Euclidean plane $(\mathbb{R}^2, \mathbb{I})$ if we form the set $\mathbb{C}\mathbb{I}X$. The latter is regular closed because $\mathbb{C}\mathbb{I}\mathbb{C}\mathbb{I}X = \mathbb{C}\mathbb{I}X$, for every X and every topological space.

In this section, we will consider several fragments of $\mathcal{S}4_u$ dealing with *regular closed sets*. From now on we will call such sets *regions*.

$\mathcal{RCC}$-8. Perhaps the best known language devised for speaking about regions is $\mathcal{RCC}$-8 which was introduced in the area of Geographical Information Systems (Egenhofer and Franzosa, 1991; Smith and Park, 1992) and as a

decidable subset of Region Connection Calculus $\mathcal{RCC}$ (Randell et al., 1992). The syntax of $\mathcal{RCC}$-8 contains *region variables* $\mathsf{r}, \mathsf{s}, \ldots$ and eight binary predicates:

- $\mathsf{DC}(\mathsf{r}, \mathsf{s})$ — regions r and s are disconnected,
- $\mathsf{EC}(\mathsf{r}, \mathsf{s})$ — r and s are externally connected,
- $\mathsf{EQ}(\mathsf{r}, \mathsf{s})$ — r and s are equal,
- $\mathsf{PO}(\mathsf{r}, \mathsf{s})$ — r and s partially overlap,
- $\mathsf{TPP}(\mathsf{r}, \mathsf{s})$ — r is a tangential proper part of s,
- $\mathsf{NTPP}(\mathsf{r}, \mathsf{s})$ — r is a nontangential proper part of s,
- the inverses of the last two—$\mathsf{TPPi}(\mathsf{r}, \mathsf{s})$ and $\mathsf{NTPPi}(\mathsf{r}, \mathsf{s})$,

which can be combined using the Boolean connectives.

The arguments of the $\mathcal{RCC}$-8 predicates, that is, region variables, are interpreted by regular closed sets—i.e., regions—of topological spaces. The following was shown in Renz, 1998 and Renz and Nebel, 1999:

THEOREM 9.3 (i) *Every satisfiable $\mathcal{RCC}$-8 formula is satisfiable in any of $\mathbb{R}^n$, for $n \geq 1$ (with region variables interpreted by connected regions only, if $n \geq 3$).*

(ii) *The satisfiability problem for $\mathcal{RCC}$-8 formulas in topological models is* NP*-complete.*

The expressive power of $\mathcal{RCC}$-8 is rather limited. It only operates with 'simple' regions and does not distinguish between connected and disconnected ones, regions with and without holes, etc. (Egenhofer and Herring, 1991). Nor can $\mathcal{RCC}$-8 represent complex relations between more than two regions. Consider, for example, three countries (say, Russia, Lithuania and Poland) such that not only each one of them is adjacent to the others, but there is a point where all the three meet (see Fig. 9.3). It can easily be shown that a ternary predicate like

$$\mathsf{EC3}(\textit{Russia}, \textit{Lithuania}, \textit{Poland}) \tag{9.13}$$

cannot be expressed in $\mathcal{RCC}$-8.

To analyse possible ways of extending $\mathcal{RCC}$-8, it will be convenient to view it as a fragment of $\mathcal{S}4_u$ (that $\mathcal{RCC}$-8 can be embedded into $\mathcal{S}4_u$ was first shown by Bennett (1994); we present here a slightly different embedding and the purpose of changes will become clear in the context of $\mathcal{BRCC}$-8 and $\mathcal{RC}$). Observe first that, for every spatial variable p, the spatial term

$$\mathbf{CI}p \tag{9.14}$$

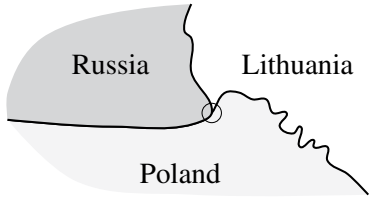


Figure 9.3. Russia, Lithuania and Poland.

is interpreted as a region (i.e., a regular closed set) in every topological model. So with every region variable r of $\mathcal{RCC}$-8 we can associate the spatial term $\varrho_{\mathsf{r}} = \mathbf{CI}p_{\mathsf{r}}$, where p_{r} is a spatial variable representing r, and then translate the $\mathcal{RCC}$-8 predicates into spatial formulas by taking

$$\begin{aligned}
\mathsf{EC}(\mathsf{r},\mathsf{s}) &= \neg(\varrho_{\mathsf{r}} \sqcap \varrho_{\mathsf{s}} = \bot) \ \wedge\ (\mathbf{I}\varrho_{\mathsf{r}} \sqcap \mathbf{I}\varrho_{\mathsf{s}} = \bot),\\
\mathsf{DC}(\mathsf{r},\mathsf{s}) &= (\varrho_{\mathsf{r}} \sqcap \varrho_{\mathsf{s}} = \bot),\\
\mathsf{EQ}(\mathsf{r},\mathsf{s}) &= (\varrho_{\mathsf{r}} \sqsubseteq \varrho_{\mathsf{s}}) \ \wedge\ (\varrho_{\mathsf{s}} \sqsubseteq \varrho_{\mathsf{r}}),\\
\mathsf{PO}(\mathsf{r},\mathsf{s}) &= \neg(\mathbf{I}\varrho_{\mathsf{r}} \sqcap \mathbf{I}\varrho_{\mathsf{s}} = \bot) \ \wedge\ \neg(\varrho_{\mathsf{r}} \sqsubseteq \varrho_{\mathsf{s}}) \ \wedge\ \neg(\varrho_{\mathsf{s}} \sqsubseteq \varrho_{\mathsf{r}}),\\
\mathsf{TPP}(\mathsf{r},\mathsf{s}) &= (\varrho_{\mathsf{r}} \sqsubseteq \varrho_{\mathsf{s}}) \ \wedge\ \neg(\varrho_{\mathsf{s}} \sqsubseteq \varrho_{\mathsf{r}}) \ \wedge\ \neg(\varrho_{\mathsf{r}} \sqsubseteq \mathbf{I}\varrho_{\mathsf{s}}),\\
\mathsf{NTPP}(\mathsf{r},\mathsf{s}) &= (\varrho_{\mathsf{r}} \sqsubseteq \mathbf{I}\varrho_{\mathsf{s}}) \ \wedge\ \neg(\varrho_{\mathsf{s}} \sqsubseteq \varrho_{\mathsf{r}})
\end{aligned}$$

(TPPi and NTPPi are the mirror images of TPP and NTPP, respectively). It should be clear that as a result we obtain the following:

THEOREM 9.4 *An $\mathcal{RCC}$-8 formula is satisfiable in a topological space iff its translation into $\mathcal{S}4_u$ defined above is satisfiable in the same topological space.*

This translation shows that in $\mathcal{RCC}$-8 any two regions can be related only in terms of truth/falsity of atomic spatial formulas of the form

$$(\varrho_1 \sqcap \varrho_2 = \bot), \qquad (\mathbf{I}\varrho_1 \sqcap \mathbf{I}\varrho_2 = \bot), \qquad (\varrho_1 \sqsubseteq \varrho_2) \quad \text{and} \quad (\varrho_1 \sqsubseteq \mathbf{I}\varrho_2),$$

where ϱ_1 and ϱ_2 are *atomic region terms*, that is, spatial terms of the form (9.14). This observation suggests two ways of increasing the expressive power of $\mathcal{RCC}$-8:

(i) by allowing the formation of *complex region terms* from atomic region terms, and

(ii) by allowing more ways of relating them (i.e., richer languages of atomic spatial formulas).

From now on we will not distinguish between a region variable r and the atomic region term ϱ_{r} representing it, and use expressions like $\mathsf{DC}(\mathsf{r},\mathsf{s})$ and $(\varrho_{\mathsf{r}} \sqcap \varrho_{\mathsf{s}} = \bot)$ as synonymous.

***BRCC*-8.** The language $\mathcal{BRCC}$-8 of Wolter and Zakharyaschev, 2000 (see also Balibiani et al., 2004) extends $\mathcal{RCC}$-8 in direction (i). It uses the same eight binary predicates as $\mathcal{RCC}$-8 and allows not only atomic regions but also their intersections, unions and complements. For instance, in $\mathcal{BRCC}$-8 we can express the fact that a region (say, the Swiss Alps) is the intersection of two other regions (Switzerland and the Alps in this case):

$$\mathsf{EQ}(\mathit{SwissAlps}, \mathit{Switzerland} \sqcap \mathit{Alps}). \tag{9.15}$$

We can embed $\mathcal{BRCC}$-8 into $\mathcal{S}4_u$ by using almost the same translation as in the case of $\mathcal{RCC}$-8. The only difference is that now, since Boolean combinations of regular closed sets are not necessarily regular closed, we should prefix compound spatial terms with **CI**. In this way we can obtain, for example, the spatial term

$$\mathbf{CI}\,(\mathit{Switzerland} \sqcap \mathit{Alps})$$

representing the Swiss Alps. In the same manner we can treat other set-theoretic operations, which leads us to the following definition of *Boolean region terms*:

$$\varrho \quad ::= \quad \mathbf{CI}p \quad | \quad \mathbf{CI}\overline{\varrho} \quad | \quad \mathbf{CI}(\varrho_1 \sqcap \varrho_2) \quad | \quad \mathbf{CI}(\varrho_1 \sqcup \varrho_2).$$

Thus $\mathcal{BRCC}$-8 can be regarded as a syntactically restricted subset of $\mathcal{S}4_u$-formulas. It follows from the above definition that Boolean region terms denote precisely the members of the well-known Boolean algebra of regular closed sets.

It is of interest to note that Boolean region terms do not increase the complexity of reasoning in arbitrary topological models: the satisfiability problem for $\mathcal{BRCC}$-8 formulas is still NP-complete. However, it becomes PSPACE-complete if all intended models are based on connected spaces ($\mathcal{BRCC}$-8 can distinguish between connected and disconnected spaces because we can express that regions r_1 and r_2 are nonempty non-tangential proper parts of a region $\mathsf{s} \neq \top$, and the union of r_1 and r_2 is precisely s:

$$\bigwedge_{i=1,2} \Big(\neg\mathsf{DC}(\mathsf{r}_i, \mathsf{r}_i) \wedge \mathsf{NTTP}(\mathsf{r}_i, \mathsf{s})\Big) \quad \wedge \quad \mathsf{NTTP}(\mathsf{s}, \mathsf{s}') \ \wedge \ \mathsf{EQ}(\mathsf{r}_1 \sqcup \mathsf{r}_2, \mathsf{s}).$$

To satisfy this formula, it suffices to take a discrete topological space with three points. But if these constraints are satisfied then both s and its complement are open and nonempty, which means that the space cannot be connected.)

On the other hand, $\mathcal{BRCC}$-8 allows some restricted comparisons of more than two regions as, e.g., in (9.15). Nevertheless, as we shall see below, ternary relations like (9.13) are still unavailable in $\mathcal{BRCC}$-8: they require different ways of comparing regions; see (ii).

***RC*.** Egenhofer and Herring (1991) proposed to relate any *two* regions in terms of the 9-intersections—3×3-matrix specifying emptiness/nonemptiness

of all (nine) possible intersections of the interiors, boundaries and exteriors of the regions. Recall that, for a region X, these three disjoint parts of the space $(U, \mathbb{I})$ can be represented as

$$\mathbb{I}X, \qquad X \cap (U - \mathbb{I}X) \qquad \text{and} \qquad U - X,$$

respectively. By generalising this approach to any finite number of regions, we obtain the fragment $\mathcal{RC}$ of $\mathcal{S}4_u$: its terms are defined as follows

$$\begin{array}{lcl} \varrho & ::= & \mathbf{CI}p \;\mid\; \mathbf{CI}\overline{\varrho} \;\mid\; \mathbf{CI}(\varrho_1 \sqcap \varrho_2) \;\mid\; \mathbf{CI}(\varrho_1 \sqcup \varrho_2), \\ \tau & ::= & \varrho \;\mid\; \mathbf{I}\varrho \;\mid\; \overline{\tau} \;\mid\; \tau_1 \sqcap \tau_2 \;\mid\; \tau_1 \sqcup \tau_2, \end{array}$$

and spatial formulas are constructed from atoms of the form $\tau_1 \sqsubseteq \tau_2$ using the Booleans (as in the full $\mathcal{S}4_u$). In other words, in $\mathcal{RC}$ we can define relations over regions in terms of inclusions of sets formed by using arbitrary set-theoretic operations on regions and their interiors. However, nested applications of the topological operators are not allowed (an example where such applications are required can be found below).

Clearly, both $\mathcal{RCC}$-8 and $\mathcal{BRCC}$-8 are fragments of $\mathcal{RC}$. Moreover, unlike $\mathcal{BRCC}$-8, the language of $\mathcal{RC}$ allows us to consider more complex relations between regions. For instance, the ternary relation required in (9.13) can now be defined as follows:

$$\begin{aligned} \mathsf{EC3}(\mathsf{r}_1, \mathsf{r}_2, \mathsf{r}_3) = \; & \neg(\varrho_{\mathsf{r}_1} \sqcap \varrho_{\mathsf{r}_2} \sqcap \varrho_{\mathsf{r}_3} = \bot) \;\wedge\; (\mathbf{I}\varrho_{\mathsf{r}_1} \sqcap \mathbf{I}\varrho_{\mathsf{r}_2} = \bot) \;\wedge \\ & (\mathbf{I}\varrho_{\mathsf{r}_2} \sqcap \mathbf{I}\varrho_{\mathsf{r}_3} = \bot) \;\wedge\; (\mathbf{I}\varrho_{\mathsf{r}_3} \sqcap \mathbf{I}\varrho_{\mathsf{r}_1} = \bot). \end{aligned}$$

Another, more abstract, example is the formula

$$\varrho_1 \sqcap \cdots \sqcap \varrho_i \;\sqcap\; \mathbf{I}\varrho'_1 \sqcap \cdots \sqcap \mathbf{I}\varrho'_j \;\sqcap\; \overline{\varrho''_1} \sqcap \cdots \sqcap \overline{\varrho''_k} \;\sqcap\; \overline{\mathbf{I}\varrho'''_1} \sqcap \cdots \sqcap \overline{\mathbf{I}\varrho'''_n} \neq \bot$$

which says that

> *regions $\varrho_1, \ldots, \varrho_i$ <u>meet</u> somewhere <u>inside</u> the region occupied jointly by all $\varrho'_1, \ldots, \varrho'_j$, but <u>outside</u> the regions $\varrho''_1, \ldots, \varrho''_k$ and <u>not inside</u> $\varrho'''_1, \ldots, \varrho'''_n$.*

Although $\mathcal{RC}$ is more expressive than both $\mathcal{RCC}$-8 and $\mathcal{BRCC}$-8, reasoning in this language is still of the same computational complexity (Gabelaia et al., 2005a):

THEOREM 9.5 *The satisfiability problem for $\mathcal{RC}$-formulas in arbitrary topological models is* NP*-complete.*

The proof follows from the fact that every satisfiable $\mathcal{RC}$-formula can be satisfied in an Aleksandrov space that is induced by a disjoint union of *n-brooms*—i.e., quasi-orders of the form depicted in Fig. 9.4. Topological spaces of this kind have a rather primitive structure satisfying the following property:

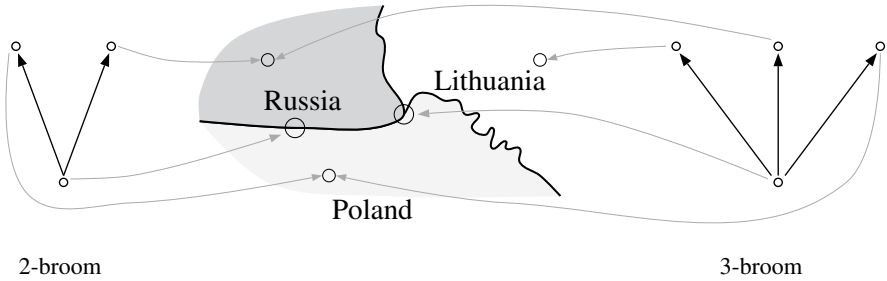


Figure 9.4. Satisfying $\mathsf{EC}(\mathit{Russia}, \mathit{Poland})$ and $\mathsf{EC3}(\mathit{Russia}, \mathit{Lithuania}, \mathit{Poland})$ in 2- and 3-brooms.

(rc) only the roots of n-brooms can be boundary points, and the minimal neighbourhood of every boundary point—i.e., the n-broom containing this point—must contain *at least one* internal point and *at least one* external point.

For example, spatial formula (9.12) cannot be satisfied in a model with this property, and so it is not in $\mathcal{RC}$.

Given a satisfiable $\mathcal{RC}$-formula φ, we can always satisfy it in a model of this kind the size of which is a polynomial (in fact, quadratic) in the length of φ, and so we have a nondeterministic polynomial time algorithm. Actually, the proof is a straightforward generalisation of the complexity proof for $\mathcal{BRCC}$-8 (Wolter and Zakharyaschev, 2000): the only difference is that in the case of $\mathcal{BRCC}$-8 it was sufficient to consider 2-brooms (which were called *forks*). This means, in particular, that ternary relation (9.13)—which is satisfiable only in a model with an n-broom, for $n \geq 3$—is indeed not expressible in $\mathcal{BRCC}$-8 (see Fig. 9.4).

REMARK 9.6 In topological terms, n-brooms are examples of so-called *door spaces* where every subset is either open or closed. However, the modal theory of n-brooms defines a wider and more interesting topological class known as *submaximal spaces* in which every dense subset is open. Submaximal spaces have been around since the early 1960s and have generated interesting and challenging problems in topology. For a survey and a systematic study of these spaces see (Arhangel'skii and Collins, 1995) and references therein.

$\mathcal{RC}^{max}$. One could go even further in direction (ii) and impose no restrictions whatsoever on the ways of relating Boolean atomic region terms. This leads us to the *maximal* fragment $\mathcal{RC}^{max}$ of $\mathcal{S}4_u$ in which spatial terms are interpreted

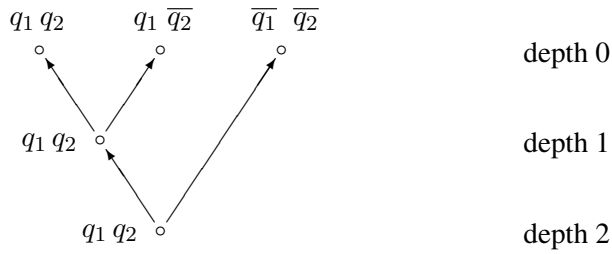


Figure 9.5. Model satisfying formula (9.16).

by regular closed sets. The syntax of its spatial terms is defined as follows:

$$\tau \quad ::= \quad \mathbf{CI}p \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \tau_1 \sqcup \tau_2 \quad | \quad \mathbf{I}\tau \quad | \quad \mathbf{C}\tau$$

and spatial formulas are constructed as in $\mathcal{S}4_u$. To understand the difference between $\mathcal{RC}^{max}$ and $\mathcal{RC}$, consider the following $\mathcal{RC}^{max}$-formula

$$(9.16) \quad \big(\mathbf{CI}q_1 \sqcap \overline{\mathbf{ICI}q_1} \neq \bot\big) \quad \wedge \quad \big(\big(\mathbf{CI}q_1 \sqcap \overline{\mathbf{ICI}q_1}\big) \quad \sqsubseteq \quad \mathbf{C}\big(\mathbf{ICI}q_1 \sqcap \mathbf{CI}q_2 \sqcap \overline{\mathbf{ICI}q_2}\big)\big).$$

It says that the boundary of $\mathbf{CI}q_1$ is not empty and that every neighbourhood of every point in this boundary contains an internal point of $\mathbf{CI}q_1$ that belongs to the boundary of $\mathbf{CI}q_2$ (compare with property (rc) above). The simplest Aleksandrov model satisfying this formula is of depth 2 (whereas n-brooms are of depth 1); it is shown in Fig. 9.5.

The price we have to pay for this expressivity is that the complexity of $\mathcal{RC}^{max}$ is the same as the complexity of full $\mathcal{S}4_u$ (Gabelaia et al., 2005a):

THEOREM 9.7 *The satisfiability problem for $\mathcal{RC}^{max}$-formulas is* PSPACE-*complete.*

This logic can also be regarded as a fragment of $\mathcal{S}4_u$ with all variables interpreted by regular closed sets.

$\mathcal{S}4_u$ with component counting. There are many ways of increasing the expressive power of $\mathcal{S}4_u$ itself. For instance, Pratt-Hartmann (2002) proposes an extension with component counting. We remind the reader that a subset X of a topological space $(U, \mathbb{I})$ is said to be *connected* if there do not exist two sets $Y_1, Y_2 \subseteq U$ such that $X \subseteq Y_1 \cup Y_2$, $X \cap Y_i \neq \emptyset$, for $i = 1, 2$, and $X \cap \mathbb{C}Y_1 \cap \mathbb{C}Y_2 = \emptyset$. Intuitively, connected sets can be thought of as consisting of 'one piece.' Then a *component* of a set X is a maximal connected subset of X. For example, the subset X of the Euclidean plane $(\mathbb{R}^2, \mathbb{I})$ in Fig. 9.2 has

only one component and so is connected, whereas its regular closure $\mathbb{CI}X$ is not connected and has two components.

The language $\mathcal{TCC}$ of Pratt-Hartmann, 2002 extends the set of atomic spatial formulas of $\mathcal{S}4_u$ with the following construct:

$$c^{\leq k}\tau,$$

where τ is a spatial term (as on p. 510) and $k \in \mathbb{N}$. The formula $c^{\leq k}\tau$ is true iff the interpretation of τ has at most k components. In particular, $c^{\leq 1}\tau$ is true iff τ is connected and $\neg c^{\leq k}\tau$ is true iff τ has at least $k+1$ components (sometimes denoted by $c^{\geq k+1}\tau$). This extension turns out to be quite expressive: for example, the $\mathcal{TCC}$-formula

$$\big(c^{\leq 1}p_1 \quad \wedge \quad c^{\leq 1}p_2 \quad \wedge \quad (p_1 \sqcap p_2 \neq \bot)\big) \quad \rightarrow \quad c^{\leq 1}(p_1 \sqcup p_2)$$

says that the union of two connected intersecting sets is also connected (here $\varphi_1 \rightarrow \varphi_2$ is an abbreviation for $\neg\varphi_1 \vee \varphi_2$). As usual, the increased expressivity results in higher complexity. The following was proved by Pratt-Hartmann (2002):

THEOREM 9.8 *The satisfiability problem for $\mathcal{TCC}$-formulas in topological models is* NEXPTIME*-complete for the binary coding of the numerical parameters.*

To conclude this section, we summarise the inclusions between the (propositional) spatial languages introduced above:

$$\mathcal{RCC}\text{-}8 \subsetneqq \mathcal{BRCC}\text{-}8 \subsetneqq \mathcal{RC} \subsetneqq \mathcal{RC}^{max} \subsetneqq \mathcal{S}4_u \subsetneqq \mathcal{TCC}.$$

3.3 Logics of distance spaces

Suppose now that we are interested in spatial logics that are capable of reasoning about spatial models based on various *distance spaces*, i.e., models of the form

$$\mathfrak{M} = (\mathfrak{D}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots), \tag{9.17}$$

where $\mathfrak{D} = (\Delta, d)$ is a distance space introduced in Sec. 3.1. If $\mathfrak{D}$ is actually a metric space then we can still use $\mathcal{S}4_u$ or its fragments interpreted in the topological space induced by $\mathfrak{D}$. However, the topological interior and closure operators $\mathbb{I}_d$ and $\mathbb{C}_d$ only deal with points that are 'infinitely close' to the given spatial object (cf. the definitions in Sec. 3.1). Being equipped with the distance function over the space, we can extend (or replace) qualitative topological reasoning by means of reasoning about distances between spatial objects. In addition to (or instead of) operators interpreted by the topological interior

and closure, we can introduce operators capable of expressing, say, that the distance from a region X to a region Y is not more than 17.

Following the 'operator-based' approach from topological logic, we arrive then to languages with 'bounded quantifiers' like $\exists^{<a}$ 'somewhere at distance $< a$' or $\forall^{<b}_{>a}$ 'everywhere within distance d for $a < d < b$,' where a and b are some numbers from $\mathbb{R}^{\geq 0}$ (or rather $\mathbb{Q}^{\geq 0}$ to avoid the problem of representing the reals).

Given a spatial model $\mathfrak{M}$ of the form (9.17), we interpret such operators in the natural way:

$$\begin{aligned}
(\exists^{<a}\tau)^{\mathfrak{M}} &= \{x \in \Delta \mid \exists y\ (d(x,y) < a \wedge y \in \tau^{\mathfrak{M}})\},\\
(\exists^{>a}\tau)^{\mathfrak{M}} &= \{x \in \Delta \mid \exists y\ (d(x,y) > a \wedge y \in \tau^{\mathfrak{M}})\},\\
(\forall^{<b}_{>a}\tau)^{\mathfrak{M}} &= \{x \in \Delta \mid \forall y\ (a < d(x,y) < b \rightarrow y \in \tau^{\mathfrak{M}})\},\\
&\text{etc.}
\end{aligned}$$

Before introducing formal languages based on these operators, it is worth having a closer look at some of them. One might be tempted to assume that the 'doughnut'-operator $\exists^{<b}_{>a}$ can be expressed via $\exists^{<b}$ and $\exists^{>a}$ by the equivalence $\exists^{<b}_{>a}\tau = \exists^{<b}\tau \sqcap \exists^{>a}\tau$. Fig. 9.6 shows that this is not the case. In the figure, we depict the regions $\exists^{<2}X$, $\exists^{>1.9}X$ and $\exists^{<2}_{>1.9}X$ for the region X consisting of the two black boxes. In particular, of all points on the plane only those in the white diamond in Fig. 9.6 (b) do not belong to $\exists^{>1.9}X$. $\exists^{<2}_{>1.9}X$ is $\exists^{<2}X$ without the three white areas in Fig. 9.6 (c). As follows from this example, $\exists^{<2}_{>1.9}X \subsetneqq \exists^{<2}X \sqcap \exists^{>1.9}X$.

In our discussion of languages for distance spaces we will formulate most results for metric spaces only. The reader is invited to consult the literature cited below to obtain detailed information about the behaviour of those languages over more general distance spaces and over Euclidean spaces.

Full 'modal' logic of distance spaces. The logic $\mathcal{MS}$ of distance spaces with the operators $\exists^{=a}$, $\exists^{<a}$, $\exists^{>a}$, $\exists^{<b}_{>a}$ (and their duals $\forall^{=a}$, $\forall^{<a}$, etc.) interpreted as defined above was introduced and analysed in (Kutz et al., 2003). Formally the *spatial terms* of this logic are defined as follows:

$$\tau \ ::= \ p_i \mid \{\ell_i\} \mid \overline{\tau} \mid \tau_1 \sqcap \tau_2 \mid \exists^{=a}\tau \mid \exists^{<a}\tau \mid \exists^{>a}\tau \mid \exists^{<b}_{>a}\tau,$$

where $a, b \in \mathbb{Q}^{\geq 0}$ with $a < b$, and the ℓ_i are *location constants* (or *nominals*) interpreted by single points, so that the $\{\ell_i\}$ are interpreted by singleton sets. As before, the *formulas* are constructed from atoms of the form $\tau_1 \sqsubseteq \tau_2$ using the Booleans ($\neg$, $\wedge$, etc.); we use $\tau_1 = \tau_2$ as an abbreviation for $\tau_1 \sqsubseteq \tau_2 \wedge \tau_2 \sqsubseteq \tau_1$.

Considering first the expressive power of $\mathcal{MS}$, one can show that over models of the form (9.17) based on metric spaces it is as expressive as the two-variable

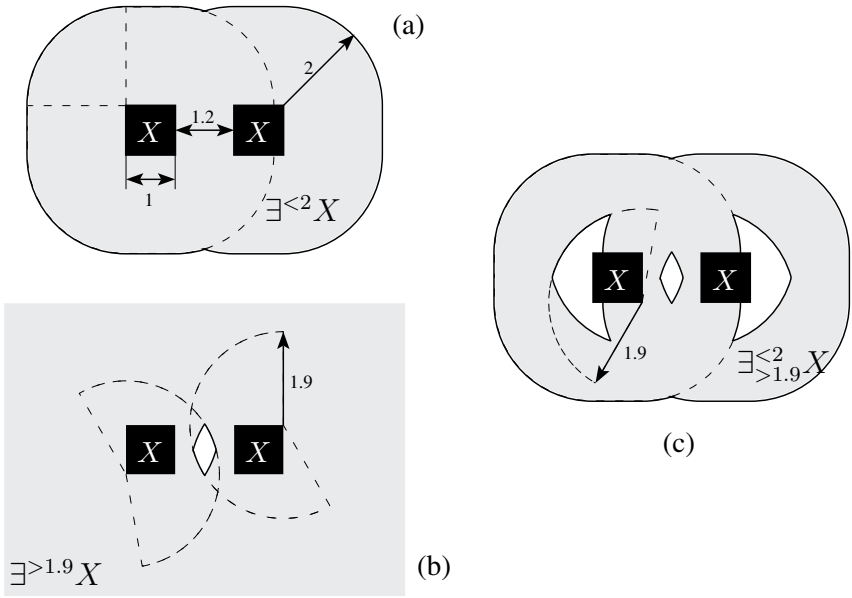


Figure 9.6. Distance operators.

fragment of first-order logic with equality, individual constants, unary predicate symbols $p_i(x)$ corresponding to spatial variables, and binary relation symbols

$$d(x, y) < a, \qquad d(x, y) = a,$$

for $a \in \mathbb{Q}^{\geq 0}$, which are interpreted in metric spaces in the obvious way (Kutz et al., 2003). Moreover, the translation between the two languages is effective.

This expressive completeness result indicates already that $\mathcal{MS}$ is indeed quite expressive. Analysing its computational properties, Kutz et al. (2003) proved that the satisfiability problem for $\mathcal{MS}$-formulas over arbitrary metric spaces is undecidable. In fact, the following much stronger theorem holds:

THEOREM 9.9 *No algorithm can decide whether an arbitrarily given $\mathcal{MS}$-formula all of whose distance operators are of the form $\exists^{<a}_{>0}$, for $a \in \mathbb{N}^{>0}$, is satisfiable in a model based on a metric space.*

The proof of this result is based on the observation that one can 'enforce' the $\mathbb{N} \times \mathbb{N}$ grid using the 'punctured' centres of circles provided by $\exists^{<a}_{>0}$.

It is worth noting that in contrast to the undecidability result above, the satisfiability problem for $\mathcal{MS}$-formulas in arbitrary distance spaces and symmetric distances spaces is *decidable*. This observation follows from the standard

translation of $\mathcal{MS}$ into the two-variable fragment of first-order logic (which is decidable in NExpTime) and the fact that reflexivity and symmetry of relations can be expressed in first-order logic using two variables only. This argument does not work for satisfiability in metric spaces because the triangle inequality cannot be expressed in first-order logic with two variables.

The logic with $\exists^{\leq a}$ and $\exists^{>a}$. Without the doughnut operators $\mathcal{MS}$ often becomes decidable and has the finite model property with respect to the intended models, that is, a formula satisfiable in a (possibly infinite) metric model is satisfiable in a finite metric model. For example, denote by $\mathcal{MS}^{\leq,>}$ the fragment of $\mathcal{MS}$ with *spatial terms* of the form

$$\tau \quad ::= \quad p_i \quad | \quad \{\ell_i\} \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \exists^{\leq a}\tau \quad | \quad \exists^{>a}\tau,$$

where $a \in \mathbb{Q}^{\geq 0}$. Kutz et al. (2003) proved that this logic has the finite model property and that the satisfiability problem for its formulas is decidable in NExpTime under the unary coding of parameters. Actually, this result was improved in Wolter and Zakharyaschev, 2005b:

Theorem 9.10 *The satisfiability problem for $\mathcal{MS}^{\leq,>}$-formulas in metric spaces is* ExpTime*-complete under the unary coding of numeric parameters in distance operators.*

The complexity of $\mathcal{MS}^{\leq,>}$-satisfiability under the binary coding of parameters remains an open research problem.

The logic with $\exists^{\leq a}$ and $\exists^{<a}$. Another interesting fragment of $\mathcal{MS}$ is based on the operators $\exists^{<a}$ and $\exists^{\leq a}$ (Wolter and Zakharyaschev, 2003). The *spatial terms* of the resulting logic $\mathcal{MS}^{\leq,<}$ are defined as follows:

$$\tau \quad ::= \quad p_i \quad | \quad \{\ell_i\} \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \exists^{\leq a}\tau \quad | \quad \exists^{<a}\tau,$$

where $a \in \mathbb{Q}^{>0}$ (by including 0 in the parameter set we would not increase the expressive power of the language, but some formulations may become awkward). The logic $\mathcal{MS}^{\leq,<}$ has the finite model property, and ExpTime-completeness can now be proved even for the binary coding of parameters:

Theorem 9.11 *The satisfiability problem for $\mathcal{MS}^{\leq,<}$-formulas in metric spaces is* ExpTime*-complete under both unary and binary coding of parameters in distance operators.*

The crucial observation in the proof of this result is that (modulo the interpretation of nominals) the logic turns out to be complete with respect to *tree metric spaces*, a feature not shared by the languages considered above. Completeness with respect to tree metric spaces makes this language also amenable to tableau-based decision procedures (Wolter and Zakharyaschev, 2003) which are not yet

available for the language $\mathcal{MS}^{\leq,>}$. An intriguing fact is that the fragments with only strict operators $\exists^{<a}$ and only non-strict ones $\exists^{\leq a}$ behave similarly in the following sense:

THEOREM 9.12 *Let φ be a formula whose only distance operators are of the form $\exists^{<a}$. Let φ' be the result of replacing occurrences of $\exists^{<a}$ in φ with $\exists^{\leq a}$. Then φ is satisfiable in a metric space iff φ' is satisfiable in a metric space.*

Of course, in the theorem above one cannot always choose the same metric space. In fact, it is worth noting that the language $\mathcal{MS}^{\leq,<}$ is properly more expressive than its fragments with only the operators $\exists^{<a}$ and $\exists^{\leq a}$, respectively. Namely, using both operators we can say that the distance between two sets p and q is precisely a:

$$\big(p \sqcap \exists^{\leq a} q \neq \bot\big) \quad \wedge \quad \big(p \sqcap \exists^{<a} q = \bot\big).$$

'Modal' logics of metric and topology. The logics of metric spaces we have considered so far can represent certain knowledge about distances between spatial objects, but are not suitable for reasoning about the induced topology. To see this for $\mathcal{MS}^{\leq,>}$ and $\mathcal{MS}^{\leq,<}$, recall that both of them have the finite model property: every satisfiable formula is satisfiable in a finite metric space. Thus, these languages cannot distinguish between finite and infinite metric spaces. On the other hand, every finite metric space induces the trivial topology in which *every set* is both closed and open. It follows that every satisfiable formula is satisfiable in a metric space with a trivial topology and that therefore the languages cannot represent anything interesting about the topology induced by a metric space. A similar argument can be used to show that $\mathcal{MS}$ itself cannot be used for representing topological knowledge.

To be able to reason about both metric and topology we can combine one of the metric logics above with one of the topo-logics considered in Sec. 3.2. Only one such combination has been investigated in detail so far: the extension of $\mathcal{S}4_u$ with the metric operators $\exists^{<a}$ and $\exists^{\leq a}$ of $\mathcal{MS}^{\leq,<}$ (Wolter and Zakharyaschev, 2005a). The terms of the resulting language we call $\mathcal{MT}$ are defined as follows:

$$\tau \quad ::= \quad p_i \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \exists^{\leq a}\tau \quad | \quad \exists^{<a}\tau \quad | \quad \mathbf{I}\tau \quad | \quad \mathbf{C}\tau,$$

where $a \in \mathbb{Q}^{>0}$. Notice that $\mathcal{MT}$ does not contain contain nominals $\{\ell_i\}$. Although it would be definitely useful to have nominals in the language, we do not include them into the signature because nothing is known about the algorithmic properties of $\mathcal{MT}$ extended by nominals. Unlike its parts $\mathcal{S}4_u$ and $\mathcal{MS}^{\leq,<}$, the logic $\mathcal{MT}$ does not have the finite model property with respect to metric spaces because the topology induced by a finite metric space is trivial. For example, the term

$$p \sqcap \mathbf{C}\overline{p}$$

is not satisfiable in any finite metric model, yet it is satisfiable in every Euclidean space.

It turns out, however, that the intended metric models for this logic can be represented in the form of relational structures (or Kripke frames), which can be regarded as partial descriptions of metric models. This representation theorem—in fact a generalisation of the McKinsey and Tarski (1944) representation theorem for topological spaces—reduces reasoning with infinite metric models to reasoning with finite relational models and can be used to show the following

THEOREM 9.13 *The satisfiability problem for $\mathcal{MT}$-formulas in metric spaces is* EXPTIME*-complete under the binary coding of parameters.*

To understand the interaction between the topological and distance operators, it is worth taking a look at the axioms required to describe this interaction. It turns out that to axiomatise the $\mathcal{MT}$-formulas that are valid in all metric models, we need the axioms governing the behaviour of the distance operators, those for the topological operators, and only two axioms where both are involved:

$$\begin{aligned} \mathbf{C}\tau &\sqsubseteq \exists^{<a}\tau, \\ \exists^{<a}\mathbf{C}\tau &\sqsubseteq \exists^{<a}\tau. \end{aligned}$$

The logic $\mathcal{MT}$ is also decidable over the real line, where it has been considered in the framework of reasoning about real-time systems (Hirshfeld and Rabinovich, 1999). It becomes undecidable, however, when we take $\mathbb{R}^2$ as the intended metric space.

Closer operator. The representation of knowledge about distances in (fragments of) $\mathcal{MS}$ is restricted to absolute distances. In particular, in $\mathcal{MS}$ it is not possible to compare distances between spatial objects without estimating the absolute values for the distances. A purely comparative approach to representing and reasoning about distance spaces would need predicates like 'X is closer to Y than it is to Z' which are quite common in our everyday life ('the body was in the middle of the room, rather closer to the door than to the window'). In the framework of spatial logics we have considered so far this predicate can be represented using the binary *closer operator* $\leftleftarrows$ with the following interpretation in distance models $\mathfrak{M} = (\mathfrak{D}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots)$:

$$(\tau_1 \leftleftarrows \tau_2)^{\mathfrak{M}} = \{x \in \Delta \mid d(x, \tau_1^{\mathfrak{M}}) < d(x, \tau_2^{\mathfrak{M}})\}. \tag{9.18}$$

In other words, $\tau_1 \leftleftarrows \tau_2$ is (interpreted by) the set containing those objects of Δ that are 'closer' (or 'more similar') to τ_1 than to τ_2. Formally, the terms of the language $\mathcal{CSL}$ of comparative distances (or similarity) are defined as follows:

$$\tau \quad ::= \quad p_i \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \tau_1 \leftleftarrows \tau_2.$$

The language $\mathcal{CSL}$ turns out to be quite powerful. Using it we can express the interior (and so the closure) operator by taking

$$\mathbf{I}\tau \;=\; \top \leftleftarrows \overline{\tau}.$$

Indeed, by the definition above, we have

$$(\mathbf{I}\tau)^{\mathfrak{M}} \;=\; \{x \in \Delta \mid d(x, \Delta - \tau^{\mathfrak{M}}) > 0\}.$$

We can also express the existential (and so the universal) modality:

$$\exists\tau \;=\; \tau \leftleftarrows \bot$$

because $d(x, \emptyset) = \infty$. Thus, $\mathcal{CSL}$ contains $\mathcal{S}4_u$ and can be regarded as a qualitative spatial formalism for reasoning about comparative distances and topology. One more interesting operator is

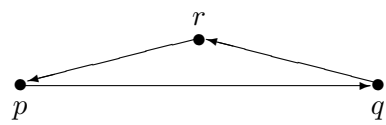
$$\tau_1 \leftrightarrows \tau_2 \;=\; \overline{(\tau_1 \leftleftarrows \tau_2)} \sqcap \overline{(\tau_2 \leftleftarrows \tau_1)}$$

which defines the set of points located at the same distance from τ_1 and τ_2.

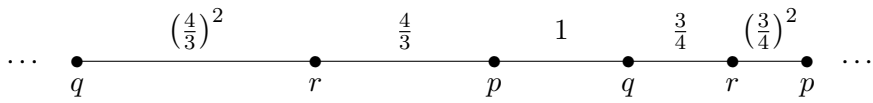
As a small illustrating example consider the formula

$$p \sqsubseteq (q \leftleftarrows r) \;\wedge\; q \sqsubseteq (r \leftleftarrows p) \;\wedge\; r \sqsubseteq (p \leftleftarrows q) \;\wedge\; p \neq \bot. \tag{9.19}$$

One can readily check that it is satisfiable in a three-point non-symmetrical model, say, in the one depicted below where the distance from x to y is the length of the shortest directed path from x to y.



On the other hand, it can be satisfied in the following subspace of $\mathbb{R}$



The following result has been obtained in Sheremet et al., 2006:

THEOREM 9.14 *The satisfiability problem for $\mathcal{CSL}$-formulas in metric spaces is* EXPTIME*-complete.*

Investigating the algorithmic properties of the combination of $\mathcal{CSL}$ with fragments of $\mathcal{MS}$ in order to facilitate reasoning about topology, comparative

distances, and absolute distances in one formalism is a challenging research problem. Define the terms of the language $\mathcal{CMS}$ by taking

$$\tau \quad ::= \quad p_i \mid \{\ell_i\} \mid \overline{\tau} \mid \tau_1 \sqcap \tau_2 \mid \exists^{\leq a}\tau \mid \exists^{<a}\tau \mid \tau_1 \leftleftarrows \tau_2,$$

where $a \in \mathbb{Q}^{>0}$. $\mathcal{CSL}$ enriched with nominals can represent *Voronoi tessellations* of various spaces. For example, let location constants ℓ_1, ℓ_2, ℓ_3 be interpreted by the points a_1, a_2, a_3 of $\mathbb{R}^2$ in Fig. 9.7. Then the $\mathcal{CMS}$-terms

$$\{\ell_i\} \leftleftarrows \{\ell_j\} \sqcup \{\ell_k\}, \quad \text{for } \{i, j, k\} = \{1, 2, 3\},$$

define the Voronoi tessellation of $\mathbb{R}^2$ corresponding to the set $\{a_1, a_2, a_3\}$.

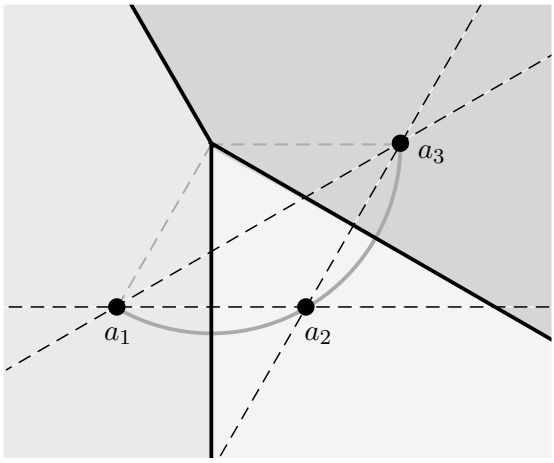


Figure 9.7. Closer operator and Voronoi tessellation.

Nothing is known about the algorithmic properties of $\mathcal{CMS}$ interpreted over arbitrary metric spaces. However, if one considers metric spaces satisfying the *min condition*

$$d(X, Y) \;=\; \min\{d(x, y) \mid x \in X, y \in Y\},$$

for all sets X and Y, then the topology induced by the metric space is trivial again and $\mathcal{CMS}$ can represent knowledge about comparative and absolute distances only (note that, by the definition, $d(X, Y) = \inf\{d(x, y) \mid x \in X, y \in Y\}$). Then we have the following result (Sheremet et al., 2005a, 2005b):

THEOREM 9.15 *The satisfiability problem for $\mathcal{CMS}$-formulas in metric spaces with the min-condition is* EXPTIME*-complete under the binary coding of parameters.*

Rather unexpectedly, over the real line $\mathbb{R}$ the logic $\mathcal{CSL}$ turns out to be undecidable, which can be proved by reduction of the (undecidable) 10th Hilbert problem on the existence of an algorithm solving arbitrary Diophantine equations; see, e.g., Barwise, 1977 and references therein. A proof can be found in Sheremet et al., 2005b.

4. Temporal logics

Now we briefly remind the reader of the two basic propositional temporal logics that will be used for speaking about the temporal dimension of spatio-temporal models introduced in Sec. 6: the linear temporal logic $\mathcal{LTL}$ and its branching time extension $\mathcal{BTL}$ (a variant of the well-known computation tree logic $\mathcal{CTL}^*$).

4.1 Linear temporal logic $\mathcal{LTL}$

Temporal logic, as opposed to first-order logic, is an approach to reasoning about time (and computation) using temporal connectives and without explicit quantification over time. Its most popular variant, the *propositional linear temporal logic* $\mathcal{LTL}$, is successfully applied in model checking as well as program verification and specification (e.g., Clarke et al., 2000; Manna and Pnueli, 1992; Manna and Pnueli, 1995).

The intended *flow of time* for $\mathcal{LTL}$ is any strict linear order $(W, <)$ with *time points* $w \in W$ and the *precedence relation* $<$. In what follows we will be mainly interested in $(\mathbb{N}, <)$ and arbitrary finite flows of time. $\mathcal{LTL}$*-formulas* are constructed from *propositional variables* $p_0, p_1, \dots$ using the Booleans and the binary *temporal operator* $\mathcal{U}$ ('until'), the intended meaning of which is as follows:

- $\varphi \,\mathcal{U}\, \psi$ stands for 'φ holds true until ψ holds.'

Other temporal connectives like $\Diamond_F$ ('sometime in the future'), $\Box_F$ ('always in the future'), and $\bigcirc$ ('at the next moment') can be defined via $\mathcal{U}$:

$$\Diamond_F \varphi = \top \,\mathcal{U}\, \varphi, \qquad \Box_F \varphi = \neg \Diamond_F \neg \varphi \qquad \text{and} \qquad \bigcirc \varphi = \bot \,\mathcal{U}\, \varphi.$$

It should be noted that we adopt the 'strict' interpretation of temporal operators, i.e., $\Box_F$, $\Diamond_F$ and $\mathcal{U}$ do not not include the present. We will use abbreviations $\Box_F^+ \varphi$ and $\Diamond_F^+ \varphi$ for $\Box_F \varphi \wedge \varphi$ and $\Diamond_F \varphi \vee \varphi$, respectively. (Note also that 'past' operators like 'since' and 'sometime in the past' can be added to the language of $\mathcal{LTL}$ as well. Here we only deal with the 'future fragment' of $\mathcal{LTL}$, as this restriction does not influence any of the results throughout.)

To evaluate $\mathcal{LTL}$-formulas in a flow of time $\mathfrak{F} = (W, <)$, we have to specify first at which time points the propositional variables hold. An $\mathcal{LTL}$*-model* is a structure of the form

$$\mathfrak{M} = (\mathfrak{F}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots),$$

where $p_i^{\mathfrak{M}} \subseteq W$ for all i. The *truth-relation* $(\mathfrak{M}, w) \models \varphi$, or simply $w \models \varphi$ if understood (which says that an $\mathcal{LTL}$-formula φ holds at moment w in $\mathfrak{M}$) is defined as follows (we omit the clauses for the Booleans):

- $w \models p_i$ iff $w \in p_i^{\mathfrak{M}}$,

- $w \models \varphi\, \mathcal{U}\, \psi$ iff there is $v > w$ such that $v \models \psi$ and $u \models \varphi$ for all $u \in (w, v)$,

where $(w, v) = \{u \in W \mid w < u < v\}$. Other temporal operators (expressible via $\mathcal{U}$) are interpreted according to their meaning. For example,

- $w \models \bigcirc\varphi$ iff $w + 1 \models \bigcirc\varphi$ (where $w + 1$ is the immediate successor of w),
- $w \models \Diamond_F\varphi$ iff there is $v > w$ such that $v \models \varphi$.

A formula φ is *satisfiable* if there is a model $\mathfrak{M}$ over $(\mathbb{N}, <)$ and a time point $n \in \mathbb{N}$ such that $(\mathfrak{M}, n) \models \varphi$. We say that φ is *finitely satisfiable* if there is a finite strict linear order $\mathfrak{F}$ and a model $\mathfrak{M}$ over it such that $(\mathfrak{M}, n) \models \varphi$ for some n in $\mathfrak{F}$.

The following results are due to Sistla and Clarke (1985):

THEOREM 9.16 *The satisfiability problem for $\mathcal{LTL}$-formulas is* PSPACE-*complete. The problem whether an $\mathcal{LTL}$-formula is finitely satisfiable is* PSPACE-*complete as well.*

This complexity result might suggest that the expressive power of $\mathcal{LTL}$ is rather limited. Surprisingly enough, this is not the case. According to the famous Kamp theorem (Kamp, 1968), the propositional temporal language with both 'until' and 'since' is as expressive as the monadic first-order language over $(\mathbb{N}, <)$ (which of course is considerably more *succinct* than $\mathcal{LTL}$).

We will also consider the fragment $\mathcal{LTL}_\Box$ of $\mathcal{LTL}$ containing only $\Box_F$ and $\Diamond_F$ as its temporal operators. The following results are due to Ono and Nakamura (1980) and Sistla and Clarke (1985):

THEOREM 9.17 *The satisfiability problem for $\mathcal{LTL}_\Box$-formulas is* NP-*complete. The problem whether an $\mathcal{LTL}_\Box$-formula is finitely satisfiable is* NP-*complete as well.*

4.2 Branching time temporal logic $\mathcal{BTL}$

The temporal logic considered above is not able to express the following statements (due to Aristotle):

- it is *necessary* that there will be a sea-battle tomorrow,
- it is *possible* that there will be a sea-battle tomorrow.

$\mathcal{LTL}$ can only say

- $\bigcirc$*sea-battle*, i.e., there will be a sea-battle tomorrow.

In other words, it does not distinguish between possible, actual, or necessary future developments. A natural way to formalise assertions of this sort is to

add two more operators A and E to the temporal language and understand them as quantifiers over 'possible histories.' For example, by interpreting E as 'it is possible that' and A as 'it is necessary that,' we can express the two Aristotle's statements by the formulas A$\bigcirc$sea-battle and E$\bigcirc$*sea-battle*, respectively.

Numerous extensions of $\mathcal{LTL}$ by means of such kind of operators have been introduced in various disciplines, in particular, computer science and artificial intelligence (Lamport, 1980; Clarke and Emerson, 1981; Emerson and Halpern, 1986) or philosophy (Prior, 1968); for more references and discussions see Thomason, 1984 and Gabbay et al., 2000. Here we only outline the essential ideas using the simple extension of $\mathcal{LTL}$ with A and E; it will be called $\mathcal{BTL}$, *branching temporal logic*.

Having fixed the language, we need to choose time structures that could allow for non-trivial interpretations. Clearly, if the flow of time is linear then at every moment the future is fixed, and so both Aφ and Eφ are equivalent to φ. The flows of time we need should be able to represent different evolutions of history. Since, on the other hand, it is natural to assume that, in contrast to the future, the past is fixed, *trees* as defined below appear to be perfect structures for modelling different histories (in particular, they correspond to the discrete tree of evolutions (9.5) of spatial transition systems).

A *tree* is a flow of time $\mathfrak{F} = (W, <)$ containing a point r, called the *root* of $\mathfrak{F}$, for which $W = \{v \mid r < v\} \cup \{r\}$, and such that for every $w \in W$, the set $\{w \mid v < w\}$ is well-founded and (strictly) linearly ordered by $<$. A *history* in $\mathfrak{F}$ is a maximal linearly $<$-ordered subset of W. Finally, an ω-tree is such a tree where every history is order isomorphic to $(\mathbb{N}, <)$.

By a *branching time model* we understand a structure

$$\mathfrak{B} = (\mathfrak{F}, \mathcal{H}, p_0^{\mathfrak{B}}, p_1^{\mathfrak{B}}, \dots),$$

where $\mathfrak{F} = (W, <)$ is an ω-tree, $\mathcal{H}$ a set of histories in $\mathfrak{F}$—the set of possible flows of time in the model—and $p_i^{\mathfrak{B}} \subseteq W$ for all i. Formulas are evaluated relative to pairs (h, w) consisting of an *actual history* $h \in \mathcal{H}$ and a time point $w \in h$. In such a pair (h, w), the temporal operators are interpreted along the actual history h as in the linear time framework, while the operators E and A quantify over the set of all histories

$$\mathcal{H}(w) = \{h' \in \mathcal{H} \mid w \in h'\}$$

coming through w. More precisely, the *truth-relation* $\models$ between models $\mathfrak{B}$ with pairs (h, w) and $\mathcal{BTL}$-formulas φ is defined inductively in the following way (we omit the clauses for the Booleans):

- $(h, w) \models p_i$ iff $w \in p_i^{\mathfrak{B}}$,
- $(h, w) \models \varphi\, \mathcal{U}\, \psi$ iff there is $v \in h$ such that $v > w$, $(h, v) \models \psi$ and $(h, u) \models \varphi$ for all $u \in (w, v)$,

- $(h, w) \models \mathsf{E}\varphi$ iff there is $h' \in \mathcal{H}(w)$ such that $(h', w) \models \varphi$,
- $(h, w) \models \mathsf{A}\varphi$ iff $(h', w) \models \varphi$ for all $h' \in \mathcal{H}(w)$.

Note that propositional variables are assumed to have no temporal aspect in the sense that their truth-values at (h, w) do not depend on the actual history h. We say that a $\mathcal{BTL}$-formula is *satisfiable* if there exists a branching time model $\mathfrak{B}$ such that $(\mathfrak{B}, h, w) \models \varphi$ for some history $h \in \mathcal{H}$ and some time point $w \in h$.

The branching time model defined above reflects the 'Ockhamist view' of time. We refer the reader to Burgess, 1979, Zanardo, 1996, Gabbay et al., 2000 and Reynolds, 2002 for more information about this and related approaches. Here we only note that our branching time logic is closely related to the computational tree logics $\mathcal{CTL}$ and $\mathcal{CTL}^*$ that are widely used in model checking and program verification and specification (Clarke and Emerson, 1981; Emerson and Halpern, 1986; Clarke et al., 2000).

It might seem more natural to quantify with E and A over the set of *all* histories in the tree rather than its subset $\mathcal{H}$. But then we would be forced to accept possibly unintended histories in $\mathfrak{F}$ as possible flows of time. Here is an example of a formula satisfiable in a branching time model as defined above, but not in a branching time model in which $\mathcal{H}$ is the set of all histories. The formula is a conjunction of the following three $\mathcal{BTL}$-formulas:

(9.20) $\mathsf{P}(\mathit{Scotland}, \mathit{UK})$,

(9.21) $\mathsf{A}\Diamond_F\Box_F\mathsf{EC}(\mathit{Scotland}, \mathit{UK})$,

(9.22) $\mathsf{A}\Box_F^+\big(\mathsf{P}(\mathit{Scotland}, \mathit{UK}) \rightarrow \mathsf{E}\bigcirc\mathsf{P}(\mathit{Scotland}, \mathit{UK})\big)$.

The first formula means that at present Scotland is part of the U.K. The second says that in all possible histories, there will be a time starting from which Scotland will be externally connected to the U.K. And the last formula claims that in all possible histories, it is always the case that if Scotland is part of the U.K. then it is still possible that it will remain in U.K. for at least one more day. (Since we do not have a combined spatio-temporal language yet, the $\mathcal{RCC}$-8 predicates $\mathsf{P}(\mathit{Scotland}, \mathit{UK})$ and $\mathsf{EC}(\mathit{Scotland}, \mathit{UK})$ should be regarded as a propositional variable and its negation, respectively.)

The following result can be obtained using a reduction to satisfiability in $\mathcal{CTL}^*$ (Hodkinson et al., 2001):

THEOREM 9.18 *The satisfiability problem for $\mathcal{BTL}$-formulas is decidable in* 2EXPTIME.

It seems that the lower bound for the computational complexity of this problem is still unknown.

REMARK 9.19 Similarly to $\mathcal{RCC}$-8, instead of time points one can take extended time entities, i.e., intervals, as primitives. This approach to temporal

representation and reasoning reflects the fact that certain assertions can be evaluated only at periods of time (e.g., 'John often drinks beer'). It was developed by Allen (1983; 1984), who observed, in particular, that relative positions of any two intervals i and j of a strict linear order can be described by precisely one of the thirteen basic interval relations: $\mathsf{before}(i,j)$, $\mathsf{meets}(i,j)$, $\mathsf{overlaps}(i,j)$, $\mathsf{during}(i,j)$, $\mathsf{starts}(i,j)$, $\mathsf{finishes}(i,j)$, their inverses ($\mathsf{before}(j,i)$, $\mathsf{meets}(j,i)$, etc.), and $\mathsf{equal}(i,j)$.

We will not consider interval temporal logics in this chapter and refer the interested reader to Vilain et al., 1989, Blackburn, 1992, Gabbay et al., 2000 and Goranko et al., 2004.

5. Combination principles

We have defined how the intended models of spatio-temporal logics (yet to be constructed) should look like. We have also identified a stock of available spatial and temporal logics to be integrated into spatio-temporal formalisms. However, we have not discussed yet how the component logics are supposed to interact with each other.

The expressive power (and consequently the computational complexity) of combined spatio-temporal formalisms obviously depends on three parameters:

- the expressiveness of the spatial component,
- the expressiveness of the temporal component, and
- the *interaction* between the two components allowed in the combined logic.

Regardless of the chosen component languages, the minimum requirement for a spatio-temporal combination to be useful is the following:

The language should be able to express changes in time of the truth-values of purely spatial propositions. **(PC)**

Languages satisfying (PC) can capture, for instance, some aspects of the *continuity of change principle* (e.g., Cohn, 1997) such as example (A) from Sec. 2: 'if two clouds are disconnected now, then at the next moment they either remain disconnected or become externally connected.' A natural way to express this principle is to encode it into the following 'spatio-temporal formula'

(A) $\mathsf{DC}(cloud_1, cloud_2) \rightarrow \bigcirc\mathsf{DC}(cloud_1, cloud_2) \vee \bigcirc\mathsf{EC}(cloud_1, cloud_2)$.

We may also need to impose some constrains on possible movements of spatial objects by comparing their positions at different moments of time. For example, the continuity principle above can be further refined by saying that the current cloud's position overlaps with its positions at the next two moments, which

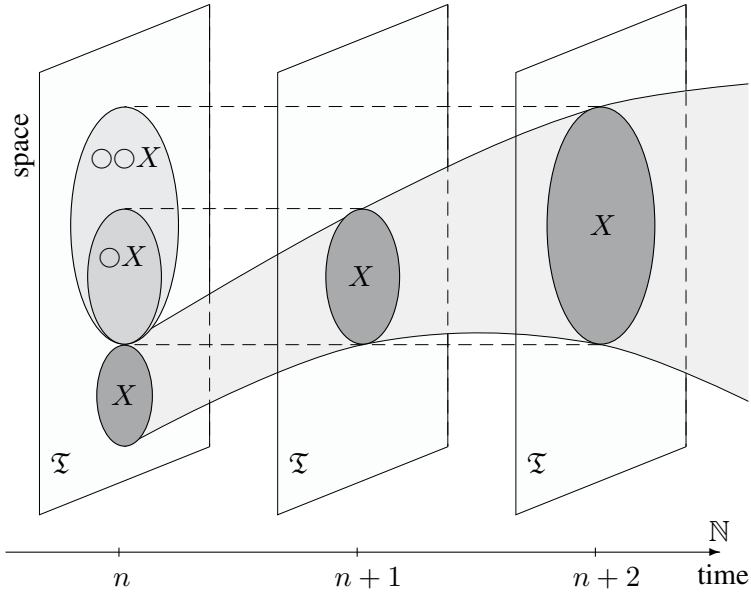


Figure 9.8. Temporal operators on regions.

requires a spatio-temporal formula of the form

(9.23) $$\mathsf{O}(\mathit{cloud}, \bigcirc \mathit{cloud}) \wedge \mathsf{O}(\mathit{cloud}, \bigcirc\bigcirc \mathit{cloud}),$$

where the predicate $\mathsf{O}(\mathsf{r}, \mathsf{s})$ means that regions r and s have at least one common interior point; it can be expressed as a disjunction of all $\mathcal{RCC}$-8 relations but DC and EC (see Fig. 9.8 where $\bigcirc X$ at moment n denotes the state of X at moment $n+1$).

The difference between (A) and (9.23) is that in the former case we apply temporal operators to spatial formulas, while in the latter to regions.

Consider now example (G) from Sec. 2: 'it will be raining over every part of England ever and ever again.' This gives rise to the formula

(G) $$\mathsf{P}(\mathit{England}, \Box_F \Diamond_F \mathit{Rain})$$

where $\mathsf{P}(\mathsf{r}, \mathsf{s}) = \mathsf{TPP}(\mathsf{r}, \mathsf{s}) \vee \mathsf{NTPP}(\mathsf{r}, \mathsf{s}) \vee \mathsf{EQ}(\mathsf{r}, \mathsf{s})$. Formula (G) can be understood as follows: all bits (points) of England will infinitely often occur in region *Rain*, but not necessarily all at the same time. Note that the formula

$$\Box_F \Diamond_F \mathsf{P}(\mathit{England}, \mathit{Rain})$$

means that it will be raining over the whole England ever and ever again.

There is an essential difference between examples (9.23) and (G). In the former, we want to control the movements of objects over a fixed finite number of steps, while in the latter example we impose restrictions on their 'asymptotic' behaviour. This leads us to two fundamental principles which will be called *local spatial object change principle* (LOC) and *asymptotic spatial object change principle* (AOC).

The language should be able to express changes or evolutions of spatial objects over some fixed finite periods of time. **(LOC)**

The language should be able to express changes or evolutions of spatial objects over the whole duration of time. **(AOC)**

In logical terms, (PC) refers to the change of truth-values of propositions, while (LOC) and (AOC) to the change of extensions of predicates.

As we shall see later on in this chapter, different combination principles result in spatio-temporal logics of different expressive power and computational complexity.

6. Combining topo-logics with temporal logics

In this section we introduce and discuss various ways of combining topo-logics and temporal logics. First we consider combinations with (fragments of) linear temporal logic $\mathcal{LTL}$ and then with branching time temporal logic $\mathcal{BTL}$.

6.1 Combinations with linear temporal logic $\mathcal{LTL}$

First we construct 'maximal' combinations with (fragments of) $\mathcal{LTL}$ meeting all three combination principles (PC), (LOC) and (AOC), and see that such a straightforward approach results in undecidable logics. Then we systematically weaken the component languages and their interaction. The result is a hierarchy of spatio-temporal logics whose complexity ranges from NP via PSPACE, EXPSPACE and 2EXPSPACE to undecidable. All omitted proofs and further details can be found in Gabelaia et al., 2005a.

As outlined in the introduction, we represent the motion of spatial objects in time using the following kind of 'snapshot' models. A *topological-temporal model* (a *tt-model*, for short) is a pair of the form $\mathfrak{M} = (\mathfrak{T}, \mathfrak{V})$, where $\mathfrak{T} = (U, \mathbb{I})$ is a topological space, and $\mathfrak{V}$, a *valuation*, is a map associating with every spatial variable p and every time point $n \in \mathbb{N}$ a set $\mathfrak{V}(p, n) \subseteq U$—the 'space' occupied by p at moment n. Such a pair $\mathfrak{M} = (\mathfrak{T}, \mathfrak{V})$ is simply a shorthand for the representation (9.3) of spatio-temporal models as a sequence of spatial models:

$$\mathfrak{M}_0 = (\mathfrak{T}, \mathfrak{V}(p_0, 0), \mathfrak{V}(p_1, 0), \dots),\ \mathfrak{M}_1 = (\mathfrak{T}, \mathfrak{V}(p_0, 1), \mathfrak{V}(p_1, 1), \dots),\ \dots$$

Combinations with (PC), (LOC) and (AOC). A 'maximalist' approach to constructing spatio-temporal logics is to allow unrestricted applications of the Booleans, the topological and the temporal operators to form spatio-temporal terms.

Denote by $\mathcal{LTL} \times \mathcal{S}4_u$ the spatio-temporal language given by the following definition:

$$\begin{aligned} \tau &::= p \mid \overline{\tau} \mid \tau_1 \sqcap \tau_2 \mid \tau_1 \sqcup \tau_2 \mid \mathbf{I}\tau \mid \mathbf{C}\tau \mid \tau_1 \,\mathcal{U}\, \tau_2, \\ \varphi &::= \tau_1 \sqsubseteq \tau_2 \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \,\mathcal{U}\, \varphi_2. \end{aligned} \tag{9.24}$$

Expressions of the form τ and φ will be called $\mathcal{LTL} \times \mathcal{S}4_u$ *terms* and *formulas*, respectively. Most of the languages we consider in this subsection are fragments of $\mathcal{LTL} \times \mathcal{S}4_u$.

As before, we can introduce the temporal operators $\Box_F$, $\Diamond_F$, and $\bigcirc$ applicable to $\mathcal{LTL} \times \mathcal{S}4_u$ formulas. Moreover, these operators can now be used to form $\mathcal{LTL} \times \mathcal{S}4_u$ terms: for example,

$$\Diamond_F \tau = \top \,\mathcal{U}\, \tau, \qquad \Box_F \tau = \overline{\Diamond_F \overline{\tau}} \qquad \text{and} \qquad \bigcirc \tau = \bot \,\mathcal{U}\, \tau,$$

where the intended meaning of $\bot$ and $\top$ is the empty set and the whole space, respectively.

$\mathcal{LTL} \times \mathcal{S}4_u$ formulas are supposed to represent propositions speaking about moving spatial objects represented by $\mathcal{LTL} \times \mathcal{S}4_u$ terms. The intended truth-values of propositions in tt-models can vary in time, but do not depend on points of spaces. But how are we to understand 'temporalised' terms?

The meaning of $\bigcirc\tau$ should be clear: at moment n, it denotes the space occupied by τ at the next moment $n+1$ (see (9.23) and Fig. 9.8). The formula

$$\mathsf{EQ}(\bigcirc\bigcirc \mathit{EU}, \mathit{EU} \sqcup \mathit{Romania} \sqcup \mathit{Bulgaria}) \tag{F}$$

formalises sentence (F) from Sec. 2. It says that in two years the EU (as it is today) will be extended with Romania and Bulgaria. Note that the formula $\bigcirc\bigcirc\mathsf{EQ}(\mathit{EU}, \mathit{EU} \sqcup \mathit{Romania} \sqcup \mathit{Bulgaria})$ has a different meaning because the EU may expand or shrink in a year. It is also not hard to formalise sentences (D), (E) and (H):

(D) $\mathsf{EQ}(\bigcirc\mathsf{X}, \mathsf{Y}) \to \neg\mathsf{EQ}(\mathsf{Y}, \bigcirc\mathsf{Y})$,

(E) $\Box_F^+\mathsf{EQ}(\bigcirc \mathit{Europe}, \mathit{Europe})$,

(H) $\Box_F^+\big(\mathsf{EQ}(\mathit{Earth}, W \sqcup L) \wedge \mathsf{EC}(W, L)\big) \wedge \mathsf{P}(W, \bigcirc W) \to \mathsf{P}(\bigcirc L, L)$.

The intended interpretation of terms of the form $\Diamond_F\tau$ and $\Box_F\tau$ is a bit more sophisticated. It reflects the standard temporal meanings of propositions '$\Diamond_F\, x \in \tau$' and '$\Box_F\, x \in \tau$,' for all points x in the topological space:

- at moment n, term $\Diamond_F\tau$ is interpreted as the union of all spatial extensions of τ at moments $m > n$;

- at moment n, term $\Box_F\tau$ is interpreted as the intersection of all spatial extensions of τ at moments $m > n$.

For example, consider Fig. 9.8 with moving cloud X depicted in it at three consecutive moments of time, and suppose X does not change after $n + 2$. Then $\Diamond_F X$ at n is the union of $\bigcirc X$ and $\bigcirc\bigcirc X$ at n and $\Box_F X$ at n is the intersection of $\bigcirc X$ and $\bigcirc\bigcirc X$ at n (i.e., $\bigcirc X$).

As another example, let us interpret the term $\Box_F\Diamond_F$*Rain* occurring in formula (G) on page 532:

- $\Diamond_F$*Rain* at moment n occupies the space where it will be raining at *some* time points $m > n$ (which may be different for different places). $\Box_F$*Rain* at n occupies the space where it will *always* be raining after n.

- $\Box_F\Diamond_F$*Rain* at n is the space where it will be raining ever and ever again after n, while $\Diamond_F\Box_F$*Rain* comprises all places where it will always be raining starting from some future moments of time.

Now, what can be the meaning of *Rain* $\mathcal{U}$ *Snow*? Similarly to the readings of $\Box_F\tau$ and $\Diamond_F\tau$ above, we adopt the following definition:

- at moment n, the spatial extension of $\tau_1\,\mathcal{U}\,\tau_2$ consists of those points x of the topological space for which there is $m > n$ such that x belongs to τ_2 at moment m and x is in τ_1 at all k whenever $n < k < m$.

The past counterpart of $\mathcal{U}$—i.e., the operator 'since' $\mathcal{S}$—can be used to say that the part of Russia that has been remaining Russian since 1917 is not connected to the part of Germany (Königsberg) that became Russian after the Second World War (Kaliningrad):

$$\mathsf{DC}(\textit{Russia}\ \mathcal{S}\ \textit{Russian_Empire}, \textit{Russia}\ \mathcal{S}\ \textit{Germany}).$$

Summing up, the valuation $\mathfrak{V}$ in tt-models can be inductively extended to arbitrary $\mathcal{LTL} \times \mathcal{S}4_u$ terms:

- $\mathfrak{V}(\overline{\tau}, n) = U - \mathfrak{V}(\tau, n), \qquad \mathfrak{V}(\tau_1 \sqcap \tau_2, n) = \mathfrak{V}(\tau_1, n) \cap \mathfrak{V}(\tau_2, n),$

- $\mathfrak{V}(\mathbf{I}\tau, n) = \mathbb{I}\,\mathfrak{V}(\tau, n), \qquad \mathfrak{V}(\mathbf{C}\tau, n) = \mathbb{C}\,\mathfrak{V}(\tau, n),$

- $\mathfrak{V}(\tau_1\,\mathcal{U}\,\tau_2, n) = \bigcup\limits_{m>n} \Big(\mathfrak{V}(\tau_2, m) \cap \bigcap\limits_{k\in(n,m)} \mathfrak{V}(\tau_1, k)\Big).$

Then we also have:

- $\mathfrak{V}(\Diamond_F\tau, n) = \bigcup\limits_{m>n} \mathfrak{V}(\tau, m), \qquad \mathfrak{V}(\Box_F\tau, n) = \bigcap\limits_{m>n} \mathfrak{V}(\tau, m),$

- $\mathfrak{V}(\bigcirc\tau, n) = \mathfrak{V}(\tau, n + 1).$

The truth-values of $\mathcal{LTL} \times \mathcal{S}4_u$ formulas in tt-models are defined as follows:

- $(\mathfrak{M}, n) \models \tau_1 \sqsubseteq \tau_2$ iff $\mathfrak{V}(\tau_1, n) \subseteq \mathfrak{V}(\tau_2, n)$,
- $(\mathfrak{M}, n) \models \neg\varphi$ iff $(\mathfrak{M}, n) \not\models \varphi$,
- $(\mathfrak{M}, n) \models \varphi_1 \wedge \varphi_2$ iff $(\mathfrak{M}, n) \models \varphi_1$ and $(\mathfrak{M}, n) \models \varphi_2$,
- $(\mathfrak{M}, n) \models \varphi_1 \,\mathcal{U}\, \varphi_2$ iff there is $m > n$ such that $(\mathfrak{M}, m) \models \varphi_2$ and $(\mathfrak{M}, k) \models \varphi_1$ for all $k \in (n, m)$.

An $\mathcal{LTL} \times \mathcal{S}4_u$ formula φ is called *satisfiable* if there exists a tt-model $\mathfrak{M}$ such that $(\mathfrak{M}, n) \models \varphi$ for some time point $n \in \mathbb{N}$.

Observe that $\mathcal{LTL} \times \mathcal{S}4_u$ contains both $\mathcal{LTL}$ and $\mathcal{S}4_u$. At first sight it may appear that the computational properties of this combination should not be too bad—after all, its spatial and temporal components are PSPACE-complete. It turns out, however, that this is very far from being the case:

THEOREM 9.20 *The satisfiability problem for $\mathcal{LTL} \times \mathcal{S}4_u$ formulas in tt-models is Σ^1_1-complete.*

It follows from Theorem 9.20 that if we strengthen the topological component to $\mathcal{TCC}$ (by allowing terms of the form $c^{\leq k}\tau$, see Sec. 3.2), then the satisfiability problem for the resulting language $\mathcal{LTL} \times \mathcal{TCC}$ is also Σ^1_1-hard. However, Theorem 9.20 is proved by a reduction of the Σ^1_1-complete *recurrent tiling problem* (Gabelaia et al., 2005b), and the terms used in its proof can denote arbitrary (i.e., not necessarily *connected*) sets. It would be interesting to know the complexity of the satisfiability problem for $\mathcal{LTL} \times \mathcal{TCC}$ formulas in tt-models where spatial variables can be interpreted at each time point by connected sets or sets containing at most k connected components for some fixed k.

One might conjecture that it is the use of the *infinitary* operators $\mathcal{U}$, $\Box_F$ and $\Diamond_F$ in the construction of $\mathcal{LTL} \times \mathcal{S}4_u$ terms that makes logics like $\mathcal{LTL} \times \mathcal{S}4_u$ 'over-expressive.' Moreover, the whole idea of tt-models based on an *infinite* flow of time may look counterintuitive in the context of spatio-temporal representation and reasoning (unlike, say, models used to represent the behaviour of reactive computer systems).

There are different approaches to avoid infinity in tt-models:

- The most radical one is to allow only *finite flows* of time. A *finite tt-model* is a triple of the form $\mathfrak{M} = (\mathfrak{T}, \mathfrak{V}, N)$, where $\mathfrak{T}$ is a topological space, $N \in \mathbb{N}$, and $\mathfrak{V}$ is a map associating with every spatial variable p and every time point $n \leq N$ a subset $\mathfrak{V}(p, n)$ of the topological space $\mathfrak{T}$.
- A more cautious approach is to impose the following *finite state assumption* on models:

FSA *Every spatial variable may have only finitely many possible states (although it may change its states infinitely often).*

Say that a (possibly infinite) tt-model $(\mathfrak{T}, \mathfrak{V})$ *satisfies* **FSA** if, for every spatial variable p, there are finitely many sets $A_1, \dots, A_m$ in the space $\mathfrak{T}$ such that $\{\mathfrak{V}(p, n) \mid n \in \mathbb{N}\} = \{A_1, \dots, A_m\}$. (Such models can be used, for instance, to capture periodic fluctuations due to season or climate changes, say, a daily tide.)

One can actually show (Gabelaia et al., 2005a) that an $\mathcal{LTL} \times \mathcal{S}4_u$ formula is satisfiable in a model with **FSA** iff it is satisfiable in a model based on a *finite* (Aleksandrov) topological space.

Unfortunately, none of these 'finitising' approaches improves the computational behaviour of the combinations too much. We can even try to weaken the temporal component to $\mathcal{LTL}_\Box$ (by allowing only the temporal operators $\Box_F$ and $\Diamond_F$ in terms and formulas), and still we have:

THEOREM 9.21 (i) *The satisfiability problem for $\mathcal{LTL}_\Box \times \mathcal{S}4_u$ formulas in (arbitrary) finite tt-models is undecidable.*

(ii) *The satisfiability problem for $\mathcal{LTL}_\Box \times \mathcal{S}4_u$ formulas in tt-models satisfying* **FSA** *is undecidable.*

However, if we weaken the spatial component further, the combinations can become decidable, with high but gradually decreasing complexity. Recall the hierarchy of topo-logics from Sec. 3.2. It suggests that next we should consider $\mathcal{RC}^{max}$ as the spatial component. In this case $\mathcal{LTL} \times \mathcal{RC}^{max}$ *terms* τ are defined by

$$\tau \ ::= \ \mathbf{CI}p \mid \overline{\tau} \mid \tau_1 \sqcap \tau_2 \mid \tau_1 \sqcup \tau_2 \mid \mathbf{I}\tau \mid \mathbf{C}\tau \mid \tau_1 \,\mathcal{U}\, \tau_2,$$

and $\mathcal{LTL} \times \mathcal{RC}^{max}$-formulas as in (9.24). Unfortunately, it is an open problem whether the satisfiability problems for $\mathcal{LTL} \times \mathcal{RC}^{max}$-formulas in arbitrary or in finite tt-models, or in tt-models satisfying **FSA** are decidable.

Let us move one more step down and denote by $\mathcal{LTL} \times \mathcal{RC}$ the language given by the following definition:

$$\begin{aligned} \varrho \ &::= \ \mathbf{CI}p \mid \mathbf{CI}\overline{\varrho} \mid \mathbf{CI}(\varrho_1 \sqcap \varrho_2) \mid \mathbf{CI}(\varrho_1 \sqcup \varrho_2) \mid \mathbf{CI}(\varrho_1 \,\mathcal{U}\, \varrho_2), \\ \tau \ &::= \ \varrho \mid \mathbf{I}\varrho \mid \overline{\tau} \mid \tau_1 \sqcap \tau_2 \mid \tau_1 \sqcup \tau_2 \end{aligned}$$

and formulas as in (9.24). Expressions of the form ϱ will be called $\mathcal{LTL} \times \mathcal{RC}$ *region terms*. Now the complexity of reasoning decreases indeed:

THEOREM 9.22 *The satisfiability problem for $\mathcal{LTL} \times \mathcal{RC}$ formulas in tt-models satisfying* **FSA**, *and in those based on (arbitrary) finite flows of time is* 2EXPSPACE*-complete.*

The existence of a 2EXPSPACE decision algorithm follows from the fact that, similarly to the case of topo-logic $\mathcal{RC}$ (without a temporal component), it

is enough to deal with *Aleksandrov* topological spaces. More precisely, it can be shown that an $\mathcal{LTL} \times \mathcal{RC}$ formula is satisfiable in a tt-model with **FSA** iff it is satisfiable in a tt-model $(\mathfrak{T}, \mathfrak{V})$ where $\mathfrak{T}$ is an Aleksandrov space induced by a finite disjoint union of finite brooms (cf. Theorem 9.5). The lower bound is established by showing that 'yardsticks' of double-exponential length (similar to those used by Stockmeyer, 1974 and Halpern and Vardi, 1989) can be encoded by $\mathcal{LTL} \times \mathcal{RC}$ formulas of polynomial length. These yardsticks can then be used to encode any Turing machine computation over double-exponential space.

By restricting the language further we obtain $\mathcal{LTL} \times \mathcal{BRCC}$-8:

$$\varphi \ ::= \ Q(\varrho_1, \varrho_2) \quad | \quad \neg\varphi \quad | \quad \varphi_1 \wedge \varphi_2 \quad | \quad \varphi_1 \vee \varphi_2 \quad | \quad \varphi_1 \, \mathcal{U} \, \varphi_2,$$

where the ϱ_i are $\mathcal{LTL} \times \mathcal{RC}$ region terms and Q ranges over (the translations of) the eight $\mathcal{RCC}$-8 predicates.

THEOREM 9.23 *The satisfiability problem for $\mathcal{LTL} \times \mathcal{BRCC}$-8 formulas in tt-models with* **FSA**, *and those that are based on* (*arbitrary*) *finite flows of time is* EXPSPACE-*complete.*

The exponential decrease in the complexity is due to the fact that now we can have a bound (linear in the size of the given formula φ) on the size of the brooms inducing the underlying Aleksandrov space of a tt-model in which an $\mathcal{LTL} \times \mathcal{BRCC}$-8 formula φ is satisfied. The lower bound can be proved by reduction of a 2^n-corridor tiling problem.

Finally, by replacing the available region terms of $\mathcal{LTL} \times \mathcal{BRCC}$-8 with

$$\varrho \quad ::= \quad \mathbf{CI}p \quad | \quad \mathbf{CI}(\varrho_1 \, \mathcal{U} \, \varrho_2)$$

we obtain the product $\mathcal{LTL} \times \mathcal{RCC}$-8. The exact complexity of the satisfiability problem for $\mathcal{LTL} \times \mathcal{RCC}$-8 formulas in tt-models satisfying **FSA**, and in (arbitrary) finite tt-models is not known. These problems are PSPACE-hard by Theorem 9.16 and in EXPSPACE by Theorem 9.23.

It is also an open problem whether satisfiability of $\mathcal{LTL} \times \mathcal{L}$ formulas in (arbitrary) tt-models is decidable, whenever $\mathcal{L} \in \{\mathcal{RCC}\text{-8}, \mathcal{BRCC}\text{-8}, \mathcal{RC}, \mathcal{RC}^{max}\}$.

Combinations with (PC) and (LOC). We can try to obtain decidable spatio-temporal combinations over *infinite* time lines by omitting the (AOC) principle and allowing only 'local control' of evolutions of spatial objects. To begin with, let us consider the fragment $\mathcal{LTL} \circ \mathcal{S}4_u$ of $\mathcal{LTL} \times \mathcal{S}4_u$ with terms of the form:

$$\tau \quad ::= \quad p \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \tau_1 \sqcup \tau_2 \quad | \quad \mathbf{I}\tau \quad | \quad \mathbf{C}\tau \quad | \quad \bigcirc\tau.$$

In other words, $\mathcal{LTL} \circ \mathcal{S}4_u$ does not allow applications of temporal operators different from $\bigcirc$ to form *terms* (but they are still available as *formula* constructors). This means that the language still satisfies (LOC), but (AOC) is no longer available.

This fragment is definitely less expressive than full $\mathcal{LTL}\times\mathcal{S}4_u$. For instance, on the one hand one can show that $\mathcal{LTL}\circ\mathcal{S}4_u$ formulas do not distinguish between arbitrary tt-models and those based on Aleksandrov topological spaces. On the other hand, the set of $\mathcal{LTL}\times\mathcal{S}4_u$ formulas satisfiable in tt-models based on Aleksandrov spaces is a proper subset of those satisfiable in arbitrary tt-models. Consider, for example, the $\mathcal{LTL}\times\mathcal{S}4_u$ formula

$$\Box_F\mathbf{I}p \;\sqsubseteq\; \mathbf{I}\Box_F p.$$

One can readily see that it is true in every tt-model based on an Aleksandrov space, but its negation can be satisfied in a tt-model. For it suffices to take the topology $\mathfrak{T}=(\mathbb{R},\mathbb{I})$ with the standard interior operator $\mathbb{I}$ on the real line, select a sequence X_n of open sets such that $\bigcap_{n\in\mathbb{N}}X_n$ is not open, e.g., $X_n=(-1/n,1/n)$, and put $\mathfrak{U}(p,n)=X_n$.

However, even this seemingly weak interaction between topological and temporal operators turns out to be dangerous:

THEOREM 9.24 *The satisfiability problem for $\mathcal{LTL}\circ\mathcal{S}4_u$ formulas in tt-models is undecidable. It is undecidable as well for tt-models satisfying* **FSA**, *and for* (*arbitrary*) *finite tt-models.*

We can try to weaken again the topological component. The language $\mathcal{LTL}\circ\mathcal{RC}^{max}$ can be obtained from $\mathcal{LTL}\times\mathcal{RC}^{max}$ by replacing the constructor $\tau_1\,\mathcal{U}\,\tau_2$ with $\bigcirc\tau$ in the definition of terms. It is not known whether this helps, that is, whether the satisfiability problem for $\mathcal{LTL}\circ\mathcal{RC}^{max}$-formulas in tt-models or in (arbitrary) finite tt-models is decidable.

If we weaken the spatial component even further, then this kind of combination turns out to be decidable. Consider the languages $\mathcal{LTL}\circ\mathcal{L}$, for $\mathcal{L}\in\{\mathcal{RC}, \mathcal{BRCC}\text{-}8, \mathcal{RCC}\text{-}8\}$, which differ from $\mathcal{LTL}\times\mathcal{L}$ only in the following aspect: in the corresponding definition of *region terms* ϱ the constructor $\mathbf{CI}(\varrho_1\,\mathcal{U}\,\varrho_2)$ is replaced with $\mathbf{CI}\bigcirc\varrho$. Then again we have a hierarchy of gradually decreasing complexity:

THEOREM 9.25 *The satisfiability problem for $\mathcal{LTL}\circ\mathcal{RC}$ formulas in tt-models is* 2EXPSPACE-*complete. It is* 2EXPSPACE-*complete as well for tt-models satisfying* **FSA** *and for* (*arbitrary*) *finite tt-models.*

THEOREM 9.26 *The satisfiability problem for $\mathcal{LTL}\circ\mathcal{BRCC}$-8 formulas in tt-models is* EXPSPACE-*complete. It is* EXPSPACE-*complete as well for tt-models satisfying* **FSA** *and for* (*arbitrary*) *finite tt-models.*

The proofs of Theorems 9.25 and 9.26 are essentially the same as those of Theorems 9.22 and 9.23. The difference is that now the correspondence between arbitrary satisfiability and satisfiability in tt-models based on Aleksandrov spaces holds not only for tt-models satisfying **FSA**, but for arbitrary tt-models as well.

THEOREM 9.27 *The satisfiability problem for $\mathcal{LTL} \circ \mathcal{RCC}$-8 formulas in tt-models is* PSPACE*-complete.*

The idea of the proof is to separate the topological and temporal parts of a given formula, and then use available satisfiability checking algorithms for the component logics (see also Theorem 9.28 below). In order to take into account the interaction between the topological and temporal parts, one has to use the so-called 'completion property' of $\mathcal{RCC}$-8 (cf. Balbiani and Condotta, 2002) with respect to a certain class $\mathfrak{C}$ of models: given a satisfiable set Φ of $\mathcal{RCC}$-8 formulas and a model in $\mathfrak{C}$ satisfying a subset of Φ, one can extend this 'partial' model to a model in $\mathfrak{C}$ satisfying the whole Φ.

The exact complexity of the satisfiability problem for $\mathcal{LTL} \circ \mathcal{RCC}$-8 formulas in tt-models satisfying **FSA**, and in (arbitrary) finite tt-models is not known. These problems are PSPACE-hard by Theorem 9.16 and in EXPSPACE by Theorem 9.26.

Combinations with (PC) only. If we want to keep the complexity low but to use an expressive topological component, then the interaction between space and time has to be weakened. One way of doing this is to consider combined languages in which the temporal operators can be applied to spatial formulas but not to spatial terms. The resulting combinations will satisfy (PC), but neither (LOC) nor (AOC) is expressible. (This way of 'temporalising' a logic was first introduced by Finger and Gabbay, 1992).

Denote by $\mathcal{LTL}[\mathcal{S}4_u]$ the spatio-temporal language given by the following definition:

$$\begin{array}{lcl} \tau & ::= & p \;\mid\; \overline{\tau} \;\mid\; \tau_1 \sqcap \tau_2 \;\mid\; \tau_1 \sqcup \tau_2 \;\mid\; \mathbf{I}\tau \;\mid\; \mathbf{C}\tau, \\ \varphi & ::= & \tau_1 \sqsubseteq \tau_2 \;\mid\; \neg\varphi \;\mid\; \varphi_1 \wedge \varphi_2 \;\mid\; \varphi_1 \vee \varphi_2 \;\mid\; \varphi_1 \,\mathcal{U}\, \varphi_2. \end{array}$$

Note that the definition of $\mathcal{LTL}[\mathcal{S}4_u]$ terms coincides with the definition of spatial terms in $\mathcal{S}4_u$ which reflects the fact that $\mathcal{LTL}[\mathcal{S}4_u]$ cannot capture (LOC) or (AOC). We have imposed no restrictions upon the temporal operators in formulas—so the combined language still contains $\mathcal{LTL}$. (Clearly, $\mathcal{S}4_u$ is a fragment of $\mathcal{LTL}[\mathcal{S}4_u]$.)

THEOREM 9.28 *The satisfiability problem for $\mathcal{LTL}[\mathcal{S}4_u]$ formulas in tt-models and in (arbitrary) finite tt-models is* PSPACE*-complete.*

The proof of this theorem is based on the fact that the interaction between spatial and temporal components of $\mathcal{LTL}[\mathcal{S}4_u]$ is rather limited. In fact, for every $\mathcal{LTL}[\mathcal{S}4_u]$ formula φ one can construct an $\mathcal{LTL}$ formula φ^* by replacing every occurrence of a (spatial) subformula $\tau_1 \sqsubseteq \tau_2$ in φ with a fresh propositional variable p_{τ_1,τ_2}. Then, given an $\mathcal{LTL}$-model $\mathfrak{M}$ for φ^* (based on $(\mathbb{N}, <)$ or a finite flow of time) and a moment n, we take the set

$$\Phi_n = \{\tau_1 \sqsubseteq \tau_2 \mid (\mathfrak{M}, n) \models p_{\tau_1,\tau_2}\} \;\cup\; \{\neg(\tau_1 \sqsubseteq \tau_2) \mid (\mathfrak{M}, n) \models \neg p_{\tau_1,\tau_2}\}$$

of spatial formulas. It is not hard to see that if Φ_n is satisfiable for every n, then there is a tt-model satisfying φ (simply because extensions of a spatial variable at different time moments are independent). Now, to check whether φ is satisfiable, it suffices to use a suitable nondeterministic algorithm (e.g., Sistla and Clarke, 1985) which guesses an $\mathcal{LTL}$-model for φ^* and then, for each time point n, to check satisfiability of Φ_n. This can be done using polynomial space in the length of φ.

Theorem 9.28 (together with Theorem 9.16) shows that the satisfiability problem for each of the spatio-temporal logics of the form $\mathcal{LTL}[\mathcal{L}]$, where $\mathcal{L} \in \{\mathcal{RCC}\text{-}8, \mathcal{BRCC}\text{-}8, \mathcal{RC}, \mathcal{RC}^{max}\}$, is also PSPACE-complete.

However, if—instead of $\mathcal{LTL}$—we consider its NP-complete fragment $\mathcal{LTL}_\Box$, the complexity of 'temporalisations' can even be lower. On the one hand, by Theorems 9.1, 9.7 and 9.28, $\mathcal{LTL}_\Box[\mathcal{S}4_u]$ and $\mathcal{LTL}_\Box[\mathcal{RC}^{max}]$ are still PSPACE-complete. On the other, by considering NP-complete topological components, the same argument as in the proof of Theorem 9.28 gives us:

THEOREM 9.29 *The satisfiability problem for $\mathcal{LTL}_\Box[\mathcal{RC}]$ formulas in tt-models is* NP*-complete.*

It follows from Theorem 9.29 that the satisfiability problems for the weaker $\mathcal{LTL}_\Box[\mathcal{RCC}\text{-}8]$ and $\mathcal{LTL}_\Box[\mathcal{BRCC}\text{-}8]$ are NP-complete as well.

6.2 Combinations with branching time temporal logic $\mathcal{BTL}$

In the framework of linear time spatio-temporal logics, we can say, for instance, that the U.K. will join the euro-zone: $\Diamond_F \mathsf{P}(\mathit{UK}, \mathit{Eurozone})$. We can also say that this will never happen. But we are not able to convey the reality, viz., that both variants are possible, that is, something like

$$\mathsf{E}\Diamond_F \mathsf{P}(\mathit{UK}, \mathit{Eurozone}) \quad \wedge \quad \mathsf{E}\neg\Box_F \mathsf{P}(\mathit{UK}, \mathit{Eurozone}). \tag{9.25}$$

In this section we summarise the results of Wolter and Zakharyaschev, 2002 on the combinations of the branching time temporal logic $\mathcal{BTL}$ with the topo-logic $\mathcal{BRCC}$-8.

The combined languages are interpreted in the following modification of tt-models. A *branching time topological model* (a *btt-model*, for short) is a quadruple $\mathfrak{M} = (\mathfrak{F}, \mathcal{H}, \mathfrak{T}, \mathfrak{V})$, where $\mathfrak{F} = (W, <)$ is an ω-tree, $\mathcal{H}$ a set of histories in $\mathfrak{F}$, $\mathfrak{T} = (U, \mathbb{I})$ a topological space, and $\mathfrak{V}$, a *valuation*, is a map associating with every spatial variable p and every time point $w \in W$ a set $\mathfrak{V}(p, w) \subseteq U$. (Observe that according to this definition, $\mathfrak{V}(p, w)$—the 'space' occupied by p at moment w—does not depend on the actual history of events.)

As concerns the languages, for each choice of topological/linear-time combination $\mathcal{L} \in \{\mathcal{LTL}[\mathcal{BRCC}\text{-}8], \mathcal{LTL} \circ \mathcal{BRCC}\text{-}8, \mathcal{LTL} \times \mathcal{BRCC}\text{-}8\}$, we have

two options: to allow applications of A and E to $\mathcal{L}$-formulas only, or to both $\mathcal{L}$-formulas and $\mathcal{L}$-region terms. The resulting languages will be denoted by $\mathcal{L}^b$ (the former option) and $\mathcal{L}^{b^x}$ (the latter one).

For example, (9.21), (9.22) and (9.25) are $\mathcal{LTL}[\mathcal{BRCC}\text{-}8]^b$-formulas. The following $(\mathcal{LTL} \circ \mathcal{BRCC}\text{-}8)^{b^x}$-formula

$$\mathsf{A}\Box^+_F\big(\mathsf{EQ}(\mathit{Europe}, \bigcirc\mathit{Europe}) \wedge \mathsf{P}(\mathit{EU}, \mathit{Europe})\big) \wedge \\ \mathsf{P}(\mathit{Europe}, \mathsf{E}\bigcirc\mathit{EU}) \wedge \mathsf{P}(\mathsf{A}\bigcirc\mathit{EU}, \mathit{EU})$$

says that, whatever happens, the region occupied by Europe will always remain the same and the EU will be part of Europe; moreover, every part of Europe has a possibility to join the EU next year, while, on the hand, what will certainly belong to the EU next year, is only part of the EU as it is today.

Now the valuation $\mathfrak{V}$ in btt-models can be inductively extended to arbitrary region terms in a way similar to the linear case: we only have to add a history as parameter. Given a region term ϱ, a history $h \in \mathcal{H}$, and a time point $w \in h$, define the *value* $\mathfrak{V}(\varrho, h, w)$ *of* ϱ *at* w *relative to* h inductively by taking

- $\mathfrak{V}(\mathbf{CI}p, h, w) = \mathbb{CI}\mathfrak{V}(p, w)$, p a spatial variable,
- $\mathfrak{V}(\mathbf{CI}\overline{\varrho}, h, w) = \mathbb{CI}(U - \mathfrak{V}(\varrho, h, w))$,
- $\mathfrak{V}(\mathbf{CI}(\varrho_1 \sqcap \varrho_2), h, w) = \mathbb{CI}(\mathfrak{V}(\varrho_1, h, w) \cap \mathfrak{V}(\varrho_2, h, w))$,
- $\mathfrak{V}(\mathbf{CI}(\varrho_1 \sqcup \varrho_2), h, w) = \mathbb{CI}(\mathfrak{V}(\varrho_1, h, w) \cup \mathfrak{V}(\varrho_2, h, w))$,
- $\mathfrak{V}(\mathbf{CI}(\varrho_1 \mathcal{U} \varrho_2), h, w) = \mathbb{CI} \bigcup\limits_{v>w,\, v\in h} \Big(\mathfrak{V}(\varrho_2, h, v) \cap \bigcap\limits_{u\in(w,v)} \mathfrak{V}(\varrho_1, h, u)\Big)$,
- $\mathfrak{V}(\mathbf{CI}\mathsf{E}\varrho, h, w) = \mathbb{CI} \bigcup\limits_{h'\in\mathcal{H}(w)} \mathfrak{V}(\varrho, h', w)$,
- $\mathfrak{V}(\mathbf{CI}\mathsf{A}\varrho, h, w) = \mathbb{CI} \bigcap\limits_{h'\in\mathcal{H}(w)} \mathfrak{V}(\varrho, h', w)$.

Now, for a formula φ and a pair (h, w), the *truth-value* of φ at (h, w) in $\mathfrak{M}$ is defined inductively as follows:

- $(\mathfrak{M}, h, w) \models Q(\varrho_1, \varrho_2)$ iff $Q(\mathfrak{V}(\varrho_1, h, w), \mathfrak{V}(\varrho_2, h, w))$ holds in $\mathfrak{T}$, for $\mathcal{RCC}$-8 predicates Q,
- $(\mathfrak{M}, h, w) \models \neg\varphi$ iff $(\mathfrak{M}, h, w) \not\models \varphi$,
- $(\mathfrak{M}, h, w) \models \varphi_1 \wedge \varphi_2$ iff $(\mathfrak{M}, h, w) \models \varphi_1$ and $(\mathfrak{M}, h, w) \models \varphi_2$,
- $(\mathfrak{M}, h, w) \models \varphi_1 \,\mathcal{U}\, \varphi_2$ iff there exists $v > w$, for $v \in h$, such that $(\mathfrak{M}, h, v) \models \varphi_2$ and $(\mathfrak{M}, h, u) \models \varphi_1$ for all $u \in (w, v)$,

- $(\mathfrak{M}, h, w) \models \mathsf{E}\varphi$ iff there is $h' \in \mathcal{H}(w)$ such that $(\mathfrak{M}, h', w) \models \varphi$,
- $(\mathfrak{M}, h, w) \models \mathsf{A}\varphi$ iff for all $h' \in \mathcal{H}(w)$, we have $(\mathfrak{M}, h', w) \models \varphi$.

A formula φ is called *satisfiable* if there is a btt-model $\mathfrak{M}$ such that $(\mathfrak{M}, h, w) \models \varphi$ for some history $h \in \mathcal{H}$ and time point $w \in h$.

THEOREM 9.30 *The satisfiability problem for* $(\mathcal{LTL} \circ \mathcal{BRCC}\text{-}8)^b$ *formulas in btt-models is decidable.*

No significant result on the computational complexity of this satisfiability problem has been obtained yet.

As to satisfiability of $\mathcal{L}^{b^x}$-formulas, we again face the problem of infinitary temporal operations on region terms. Now, besides the linear temporal operators, the region terms can also be affected by the 'branch' operators A and E. In fact, at least for *discrete* topological spaces (i.e., spaces $\mathfrak{T} = (U, \mathbb{I})$ in which $\mathbb{I}$ is the identity function) we have the following negative result:

THEOREM 9.31 *The satisfiability problem for* $(\mathcal{LTL} \times \mathcal{BRCC}\text{-}8)^{b^x}$ *formulas in btt-models based on discrete topological spaces is undecidable.*

We conjecture that the satisfiability problem for $(\mathcal{LTL} \times \mathcal{BRCC}\text{-}8)^{b^x}$ formulas in btt-models based on arbitrary topological and Euclidean spaces is undecidable as well.

A natural way to search for decidable variants of the undecidable logics discussed above is to restrict the class of btt-models to those having finite sets of histories and where each history satisfies the finite state assumption. We conjecture that the satisfiability problem for $(\mathcal{LTL} \times \mathcal{BRCC}\text{-}8)^{b^x}$ formulas in this kind of btt-models is decidable.

REMARK 9.32 Temporalisations of $\mathcal{RCC}$-8 and $\mathcal{BRCC}$-8 with the help of Allen's interval calculus (see Remark 9.19) were considered by Bennett et al., 2002, Gerevini and Nebel, 2002 and Gabbay et al., 2003.

7. Combining distance logics with temporal logics

Unfortunately, not so much is known about temporal extensions of logics of distance spaces. Of course, some of the 'negative' results from Sec. 6 hold for similar combinations with those distance logics that contain $\mathcal{S}4_u$ as a sub-logic (for example, $\mathcal{MT}$ or $\mathcal{CSL}$). The technique of the proof of Theorem 9.28 can be used to show that the temporalisations $\mathcal{LTL}[L]$ of the logics L of distance spaces from Sec. 3.3 (where the temporal operators can only be applied to formulas but not to spatial terms) inherit the complexity of L (see the end of this section). And finally, some of the methods developed to deal with products of modal (in particular, temporal) logics (see Gabbay et al., 2003 and references

therein) can be applied to analyse the computational behaviour of combinations of $\mathcal{LTL}$ with distance logics like $\mathcal{MS}^{\leq}$ which only contains distance operators of the form $\exists^{\leq a}$ (and their duals).

Denote by $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ the spatio-temporal language satisfying the (PC), (LOC) and (AOC) principles and given by the following definition:

$$\begin{array}{lllllllllll} \tau & ::= & p_i & | & \overline{\tau} & | & \tau_1 \sqcap \tau_2 & | & \tau_1 \sqcup \tau_2 & | & \exists^{\leq a}\tau & | & \tau_1 \,\mathcal{U}\, \tau_2, \\ \varphi & ::= & \tau_1 \sqsubseteq \tau_2 & | & \neg\varphi & | & \varphi_1 \wedge \varphi_2 & | & \varphi_1 \vee \varphi_2 & | & \varphi_1 \,\mathcal{U}\, \varphi_2. \end{array}$$

As before, expressions of the form τ and φ are called $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ *terms* and *formulas*, respectively.

A *metric temporal model* (*mt-model*, for short) is a pair of the form $\mathfrak{M} = (\mathfrak{D}, \mathfrak{V})$, where $\mathfrak{D} = (\Delta, d)$ is a metric space and $\mathfrak{V}$, a valuation, is a map associating with each spatial variable p and each time instant n a set $\mathfrak{V}(p, n) \subseteq \Delta$. The valuation can be inductively extended to arbitrary $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ terms in a straightforward way:

- $\mathfrak{V}(\overline{\tau}, n) = \Delta - \mathfrak{V}(\tau, n), \qquad \mathfrak{V}(\tau_1 \sqcap \tau_2, n) = \mathfrak{V}(\tau_1, n) \cap \mathfrak{V}(\tau_2, n),$
- $\mathfrak{V}(\exists^{\leq a}\tau, n) = \big\{x \in \Delta \mid \exists y(d(x, y) < a \wedge y \in \mathfrak{V}(\tau, n))\big\},$
- $\mathfrak{V}(\tau_1 \,\mathcal{U}\, \tau_2, n) = \bigcup\limits_{m>n} \Big(\mathfrak{V}(\tau_2, m) \cap \bigcap\limits_{k \in (n,m)} \mathfrak{V}(\tau_1, k)\Big).$

The truth-values of $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ formulas in mt-models are defined in precisely the same way as for spatio-temporal logics from Sec. 6.1. As before, we freely use the temporal operators $\bigcirc$, $\Diamond_F$ and $\Box_F$ (as well as their non-strict versions $\Diamond_F^+$ and $\Box_F^+$).

As an example of an $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ formula, consider the following formalisation of (I) from Sec. 2:

$$\bigwedge_{i=1,2} \big((\mathit{desert}_i \neq \bot) \;\wedge\; \Box_F^+\big(\exists^{\leq a}\mathit{desert}_i \sqsubseteq \bigcirc\mathit{desert}_i\big)\big) \;\rightarrow \\ \Diamond_F \Box_F \big(\mathit{desert}_1 \sqcap \mathit{desert}_2 \neq \bot\big).$$

It says that two nonempty deserts (say, the Kalahari and the Sahara) increasing their size in all directions by at least some $a \in \mathbb{Q}^{>0}$ each year will eventually intersect. Notice that this formula is valid in mt-models based on Euclidean spaces, but not in models based on disconnected or discrete metric spaces.

Unfortunately, the complexity of this combination of a PSPACE-complete and an EXPTIME-complete logics turns out to be too high. Using an almost straightforward encoding of the recurring tiling problem (see, e.g., the proof of Theorem 11.1 in Gabbay et al., 2003) one can prove the following:

THEOREM 9.33 *The satisfiability problem for $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ formulas in mt-models is Σ^1_1-complete.*

This result might suggest that combinations of $\mathcal{LTL}$ with $\mathcal{MS}^{\leq}$ have the same computational properties as combinations with $\mathcal{S}4_u$ in Sec. 6.1. However, this is not the case. To see the difference, let us consider the problem of *term satisfiability* for both languages: a term τ is *satisfiable* if there is a model $\mathfrak{M}$ for the language where $\mathfrak{V}(\tau, n)$ is not empty for some time moment n. It can be shown (similarly to the proof of Theorem 9.20) that the satisfiability problem for $\mathcal{LTL} \times \mathcal{S}4_u$ terms is Σ^1_1-complete. In the case of $\mathcal{MS}^{\leq}$ the picture is slightly better—the problem is decidable but not in time bounded by any 'tower' of exponents:

THEOREM 9.34 *The satisfiability problem for $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ terms in mt-models is decidable, but not in elementary time.*

The proof requires three ingredients. First, one can show (similarly to the proof of Theorem 9.11) that any satisfiable $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ term is satisfiable in an mt-model based on a *tree metric space*. This observation makes it possible to apply the methods developed to analyse the product modal logic $\mathbf{PTL}_{\Box\bigcirc} \times \mathbf{K}$. In particular, the decidability result is proved analogously to Theorem 13.6 from Gabbay et al., 2003: first, mt-models are represented in the form of *quasimodels*, and then the existence of a quasimodel for a given term is encoded in monadic second-order logic. The non-elementary lower bound can be established by a polynomial reduction of the satisfiability problem for $\mathbf{PTL}_{\Box\bigcirc} \times \mathbf{K}$ (which is non-elementary by Theorem 6.37 and Claim 6.25 of Gabbay et al., 2003) to satisfiability of $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ terms.

It is worth noting that the language of $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ terms is 'local' in the sense that every term refers to a bounded area of the metric space, and the size of this area can be effectively computed. (In particular, statement (I) refers to the whole space, and so cannot be expressed in the language of $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ terms.) In fact, this is the crucial observation required for the decidability result. $\mathcal{LTL} \times \mathcal{MS}^{\leq}$ formulas, on the contrary, can speak about the whole space which makes it possible to simulate tilings.

In the same way one can explain the computational behaviour of the combination $\mathcal{LTL} \circ \mathcal{MS}^{\leq}$ satisfying both (PC) and (LOC), but not (AOC). It is defined analogously to the spatio-temporal case by replacing $\tau_1 \,\mathcal{U}\, \tau_2$ with $\bigcirc\tau$ in the definition of terms:

$$\tau \quad ::= \quad p_i \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \tau_1 \sqcup \tau_2 \quad | \quad \exists^{\leq a}\tau \quad | \quad \bigcirc\tau.$$

The formulas of $\mathcal{LTL} \circ \mathcal{MS}^{\leq}$ are defined in the same way as above.

As Theorem 9.24 might suggest, the same negative result holds even for this restricted combination:

THEOREM 9.35 *The satisfiability problem for $\mathcal{LTL} \circ \mathcal{MS}^{\leq}$ formulas in mt-models is undecidable.*

On the other hand, $\mathcal{LTL} \circ \mathcal{MS}^{\leq}$ terms are basically 'harmless' because they can only speak about limited time, at most $\ell(\tau)$ time moments, to be more precise (where $\ell(\tau)$ is the length of τ). So we have the following:

THEOREM 9.36 *The satisfiability problem for $\mathcal{LTL} \circ \mathcal{MS}^{\leq}$ terms in mt-models is* EXPTIME*-complete under both unary and binary coding of parameters in distance operators.*

Finally, the temporalisations of distance logics satisfying only the (PC) principle, and so containing no temporal operators in spatial terms, inherit the higher complexity of the spatial component, which is proved similarly to the proof of Theorem 9.28:

THEOREM 9.37 *The satisfiability problem for $\mathcal{LTL}[\mathcal{MS}^{\leq,<}]$, $\mathcal{LTL}[\mathcal{MT}]$, and $\mathcal{LTL}[\mathcal{CMS}]$ formulas in mt-models is* EXPTIME*-complete for both unary and binary coding of parameters.*

8. Logics for dynamical systems

The snapshot models and the corresponding spatio-temporal logics discussed above are a convenient tool for representing and reasoning about evolutions of spatial configurations of regions such as the political (geographical, weather, etc.) map of the changing world, where we are interested in keeping track of the relations between regions.

On the other hand, if we want to model how an object moves over an otherwise stable space and keep track of its asymptotic trajectory then different models of space and time may be preferable, namely models corresponding to dynamical systems (e.g., Brown, 1976; Katok and Hasselblatt, 1995).

A *dynamical model* is a pair of the form

$$\mathfrak{A} = (\mathfrak{M}, g), \tag{9.26}$$

where $\mathfrak{M} = (\mathfrak{S}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots)$ is a spatial model and g is a total function on the space $\mathfrak{S}$. Often g is required to satisfy certain constraints depending on the structure of $\mathfrak{S}$. For example, if $\mathfrak{S}$ is a topological space, then g is often required to be continuous or even a bijective continuous and open mapping (that is, a homeomorphism).

In the framework of such models, we are interested in the *orbits*

$$Orb_g(w) = \{g(w),\ g^2(w),\ \dots\}$$

of certain points w from $\mathfrak{S}$ (representing moving objects). The model $\mathfrak{M}$ describes a spatial environment in which w moves according to the rule (law)

g. A typical question in the framework of dynamical models is whether a point from a region $p_0^{\mathfrak{M}}$ will eventually reach $p_1^{\mathfrak{M}}$ without visiting $p_2^{\mathfrak{M}}$, or whether the rule g is such that w will be returning to $p_1^{\mathfrak{M}}$ infinitely often.

The aim of this section is to discuss the existing logics capable of talking about some aspects of dynamical models. In particular, we consider the relation between the spatio-temporal logics above and logics for dynamical models. Before reading this section the reader is recommended to have a look at Ch. 10.

Let us begin with two illuminative examples.

A physical system. Consider a physical system with a single degree of freedom, say, a body having mass m and moving along some axis. The movement of the body in a force field $f(x,t)$ can be described by the following system of differential equations:

$$\begin{aligned} \dot{x}(t) &= v(t), \\ \dot{v}(t) &= f(x,t)/m, \end{aligned}$$

where $x(t)$ and $v(t)$ are, respectively, the position and the velocity of the body at time t. For every initial point $(x_0, v_0) \in \mathbb{R}^2$, the differential equations determine the trajectory $\pi_{(x_0,v_0)}(t)$ of the body (more precisely, its position and velocity) that starts with the velocity v_0 at the position x_0 and moves according to the above equations. The collection of all those trajectories for different initial conditions form the *phase portrait* of the differential equation (depicted in the left-hand side of Fig. 9.9).

Now consider the function $\phi((x,v),t)$, called the *flow* of the equations, defined by taking $\phi((x,v),t) = \pi_{(x,v)}(t)$. Note that

- $\phi((x,v),0) = (x,v)$ and
- $\phi((x,v),t+s) = \phi(\phi((x,v),t),s)$.

The graph of this function represents trajectories in $\mathbb{R}^2 \times \mathbb{R}$ and the phase portrait can be considered as the projection of $\phi((x,v),t)$ onto $\mathbb{R}^2$; see Fig. 9.9. Note also that $\phi((x,v),t)$ is continuous in all coordinates.

Given such a physical system, we usually want to know answers to the following standard questions. Suppose that the initial conditions (position and velocity) of the body are restricted by some set $I \subseteq \mathbb{R}^2$. Is it the case that starting from any point of I the body will eventually reach some point in another set F? Will it be visiting F infinitely often? Is it the case that the body will never hit some 'danger zone' $D \subseteq \mathbb{R}^2$?

A dynamical model $(\mathfrak{M}, g)$ for the differential equations above can be defined as follows. The underlying space $\mathfrak{S}$ of $\mathfrak{M}$ is the Euclidean plane $\mathbb{R}^2$. Let $g(x,v) = \phi((x,v),\delta)$, for some fixed small time unit $\delta > 0$. The predicates $p_i^{\mathfrak{M}}$ can model the initial and final conditions I and F, the danger zone D,

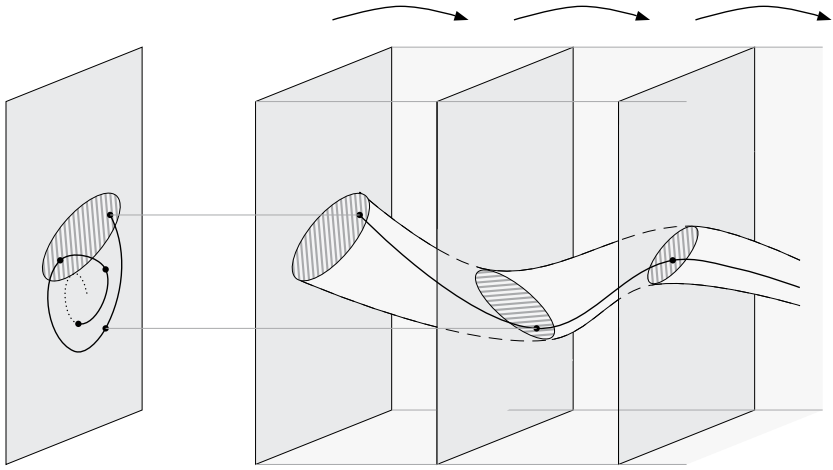


Figure 9.9. Dynamical system.

etc. As g is easily seen to be continuous, $(\mathfrak{M}, g)$ is a dynamical model with a continuous function on $\mathbb{R}^2$. The three questions above can then be formalised as whether we have

- $I \subseteq \bigcup_{i>0} g^{-i}(F)$,
- $I \subseteq \bigcap_{j>0} g^{-j}\big(\bigcup_{i>0} g^{-i}(F)\big)$,
- $I \cap \bigcup_{i>0} g^{-i}(D) = \emptyset$.

It is to be noted that, on the other hand, the flow $\phi((x, v), t)$ can be regarded as a snapshot spatio-temporal model

$$\mathfrak{M}_0, \ \mathfrak{M}_1, \ \ldots,$$

where $\mathfrak{M}_i = \big(\mathfrak{S}, g^{-i}(p_0^{\mathfrak{M}}), g^{-i}(p_1^{\mathfrak{M}}), \ldots\big)$, for $i \geq 0$. The intuition behind this definition is as follows: a point (x, v) belongs to a set $Y \subseteq \mathbb{R}^2$ at time point i iff (x, v) is moved to Y by i consecutive applications of g, that is, $(x, v) \in g^{-i}(Y)$.

Game of Life. Our second example is the *Game of Life* invented by J.H. Conway in the 1970s (e.g., Allouche et al., 2001). The game is defined as follows. We have a finite $\{1, \ldots, n\} \times \{1, \ldots, n\}$ or an infinite $\mathbb{Z} \times \mathbb{Z}$ board.

Each point on the board is either occupied or vacant (living or dead). At each regular time step the points of the board simultaneously change according to the following rules:

(birth) a vacant point with exactly three occupied neighbours becomes an occupied cell,

(survival) an occupied point with two or three occupied neighbours stays occupied,

(death) in all other cases, the point becomes or remains vacant.

Thus, at each step $i \geq 0$ the state of the game can be represented by the spatial model

$$\mathfrak{M}_i = (\mathfrak{S}, o^{\mathfrak{M}_i}, v^{\mathfrak{M}_i}), \tag{9.27}$$

where $\mathfrak{S}$ is the board, $o^{\mathfrak{M}_i}$ is the set of occupied points at step i and $v^{\mathfrak{M}_i}$ the set of vacant ones.

The Game of Life can be represented by the spatial transition system which consists of all possible models $\mathfrak{M}$ of the form (9.27), and $\mathfrak{M} \rightarrowtail \mathfrak{M}'$ holds iff $\mathfrak{M}'$ is obtained from $\mathfrak{M}$ by one step of the game according to the rules above. As the Game of Life is deterministic (for every $\mathfrak{M}$ there is exactly one $\mathfrak{M}'$ such that $\mathfrak{M} \rightarrowtail \mathfrak{M}'$), there is exactly one evolution for any spatial transition system representing it. In other words, for every initial state of the Game we obtain exactly one snapshot model.

The Game of Life (on, say, $\mathbb{Z} \times \mathbb{Z}$) can also be formalised as a dynamical model

$$\mathfrak{N} = \big((\mathfrak{T}, p_0^{\mathfrak{N}}, p_1^{\mathfrak{N}}, \dots), g\big).$$

The underlying space $\mathfrak{T}$ is comprised of all functions from $\mathbb{Z} \times \mathbb{Z}$ into $\{o, v\}$ representing distributions of occupied and vacant points, that is, $\mathfrak{T} = \{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$. The function g maps every $\eta \in \{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$ to the function $g(\eta) \in \{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$ representing the next distribution of occupied and vacant points. In other words, the underlying space can be regarded as the set of all models $(\mathfrak{S}, o^{\mathfrak{M}}, v^{\mathfrak{M}})$ with the function g given by the transition relation (rule) $\rightarrowtail$. Finally, define a metric d on $\{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$ so that g becomes a continuous function for the induced topology $\mathbb{I}_d$ as follows. Set, for $\eta_1, \eta_2 \in \{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$,

$$d(\eta_1, \eta_2) = \frac{1}{k}$$

iff η_1 and η_2 agree on all points within the $k \times k$ square

$$I_k = \{(n, m) \in \mathbb{Z} \times \mathbb{Z} \mid \max\{n, m\} < k\}$$

but disagree on at least one point in I_{k+1}. One can show that the metric d defines a compact topological space on $\{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$ with respect to which g is continuous.

Notice that in this dynamical model predicates are not subsets of the board $\mathbb{Z} \times \mathbb{Z}$ but of $\{o, v\}^{\mathbb{Z}\times\mathbb{Z}}$. We can take, for instance, some interesting set $p_0^{\mathfrak{N}}$ of initial states (i.e., models $(\mathfrak{S}, o^{\mathfrak{M}}, v^{\mathfrak{M}})$), say, those with precisely N living points, and check whether all of them (or at least one of them) will eventually 'die out,' that is, reach the singleton set $p_1^{\mathfrak{N}} = \{(\mathfrak{S}, o^{\mathfrak{M}}, v^{\mathfrak{M}})\}$ where $o^{\mathfrak{M}}$ is empty.

The resulting dynamical model $\mathfrak{N} = \big((\mathfrak{T}, p_0^{\mathfrak{N}}, p_1^{\mathfrak{N}}, \dots), g\big)$ can be 'unravelled' into the transition system $s_0 \rightarrowtail s_1 \rightarrowtail \dots$ where

$$\mu(s_n) = \big(\mathfrak{T}, g^{-n}(p_0^{\mathfrak{N}}), g^{-n}(p_1^{\mathfrak{N}}), \dots\big).$$

8.1 Dynamic topological logics

We start our discussion of languages for reasoning about dynamical systems by considering dynamical models based on various topological spaces.

A *dynamic topological model* (DTM, for short) is a pair

$$\mathfrak{A} = (\mathfrak{M}, g),$$

where $\mathfrak{M} = (\mathfrak{T}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots)$ is a topological model based on a topological space $\mathfrak{T} = (U, \mathbb{I})$ and g is a function on $\mathfrak{T}$. The minimum requirement imposed on g in dynamical systems is its continuity. We remind the reader that a function g on $\mathfrak{T}$ is called *continuous* if $g^{-1}(X)$ is open whenever $X \subseteq U$ is open. If $g(X)$ is open whenever X is open, then g is called *open*. Another important type of functions is *homeomorphisms*, that is, bijective continuous and open functions on $\mathfrak{T}$. (It is also usually assumed that the underlying topological spaces are compact. We will not make this assumption in general, but point out when our results hold for compact topological spaces.)

The language we consider for representing and reasoning about dynamic topological systems is slightly different from most of the languages for snapshot models because, as we have already seen, in dynamical systems we are more interested in following the orbit of an object in space and time rather than in comparing the relative positions of regions in space. That is why the language $\mathcal{DTL}$ for reasoning about topological systems is 'local' in the sense that we see the space from the windows of our moving 'car' as opposed to the 'global' language of spatio-temporal logics from Sec. 6.1 where we could observe all moving 'cars' and their relative positions. Formally, this means that we represent knowledge about the evolution of objects by means of terms and do not consider formulas constructed from them.

The set of $\mathcal{DTL}$*-terms* τ is defined as follows:

$$\tau \quad ::= \quad p \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \mathbf{I}\tau \quad | \quad \bigcirc\tau \quad | \quad \Box_F\tau \quad | \quad \Diamond_F\tau.$$

It is worth noting that the addition of the operator $\mathcal{U}$ for 'until' to the set of constructors for terms would not affect any of the results presented below. We have omitted 'until' to keep the language as simple as possible.

In a dynamic topological model $\mathfrak{A} = (\mathfrak{M}, g)$, terms τ are interpreted as sets $\tau^{\mathfrak{A}} \subseteq U$, where $\mathfrak{M}$ is based on the topological space $(U, \mathbb{I})$. Clearly, $p_i^{\mathfrak{A}} = p_i^{\mathfrak{M}}$ for every spatial variable p_i. The Boolean operators and the operator **I** are interpreted as before. The interpretation of the temporal operators on terms should become clear from the following consideration: for a point $w \in U$ and a term τ, we have

$$w \in (\bigcirc\tau)^{\mathfrak{A}} \quad \text{iff} \quad g(w) \in \tau^{\mathfrak{A}} \quad \text{iff} \quad w \in g^{-1}(\tau^{\mathfrak{A}}). \tag{9.28}$$

Roughly, a time point n in a snapshot model corresponds to n applications of the function g. If we understand $w \in (\Diamond_F\tau)^{\mathfrak{A}}$ as 'eventually w will be moved by g to $\tau^{\mathfrak{A}}$' and $w \in (\Box_F\tau)^{\mathfrak{A}}$ as 'g will always keep w in $\tau^{\mathfrak{A}}$,' then

$$w \in (\Diamond_F\tau)^{\mathfrak{A}} \quad \text{iff} \quad Orb_g(w) \cap \tau^{\mathfrak{A}} \neq \emptyset \quad \text{iff} \quad w \in \bigcup_{i>0} g^{-i}(\tau^{\mathfrak{A}}), \tag{9.29}$$

$$w \in (\Box_F\tau)^{\mathfrak{A}} \quad \text{iff} \quad Orb_g(w) \subseteq \tau^{\mathfrak{A}} \quad \text{iff} \quad w \in \bigcap_{i>0} g^{-i}(\tau^{\mathfrak{A}}). \tag{9.30}$$

For example, $w \in \big(p_1 \sqcap \Diamond_F p_2\big)^{\mathfrak{A}}$ means that w is in $p_1^{\mathfrak{A}}$ and reaches $p_2^{\mathfrak{A}}$ by a finite number of iterations of g.

In this section, we are interested in the satisfiability and validity problem for $\mathcal{DTL}$-terms in some important classes of dynamic topological models:

- A $\mathcal{DTL}$-term τ is *satisfiable* in a class $\mathcal{M}$ of DTMs iff there exists $\mathfrak{A} \in \mathcal{M}$ such that $\tau^{\mathfrak{A}} \neq \emptyset$.
- A $\mathcal{DTL}$-term τ is *valid* in a class $\mathcal{M}$ of DTMs iff $\overline{\tau}$ is not satisfiable in $\mathcal{M}$—i.e., iff $\tau^{\mathfrak{A}}$ coincides with the whole space for every $\mathfrak{A} \in \mathcal{M}$.

DTMs with homeomorphisms. We first connect satisfiability in certain dynamic topological models with satisfiability in snapshot topological temporal models. The discussion of the two examples above indicates already how one can go back and forth between snapshot topological models and dynamic topological models. More precisely, one can show the following:

THEOREM 9.38 *Let $\mathcal{M}$ be any of the following classes of dynamic topological models:*

- *DTMs based on Aleksandrov spaces with homeomorphisms;*
- *DTMs based on topological spaces with homeomorphisms;*
- *DTMs based on $\mathbb{R}^n$ with homeomorphisms, for $n > 1$;*

- *DTMs based on the n-dimensional unit ball with a measure preserving homeomorphism, for $n > 1$.*

Then a $\mathcal{DTL}$-term τ is satisfiable in $\mathcal{M}$ iff the formula $\neg(\tau = \bot)$ is satisfiable in a snapshot tt-model based on a topological space underlying some model from the class $\mathcal{M}$.

It should not come as a surprise now that reasoning with $\mathcal{DTL}$-terms about these classes of DTMs can be extremely complex. The following result was proved in Konev et al., 2006b by reduction of Post's correspondence problem:

THEOREM 9.39 *Let $\mathcal{M}$ be any of the classes of DTMs mentioned in Theorem 9.38. Then the set of $\mathcal{DTL}$-terms that are valid in models from $\mathcal{M}$ is not recursively enumerable.*

It is worth noting that the four sets of terms that are valid in the classes of models mentioned in Theorem 9.38 are all different. As was shown by Slavnov (2003), the term

$$\mathbf{I}\Diamond_F\big(p \sqcap \mathbf{CI}\overline{p}\big)$$

is not satisfiable in any DTM based on $(\mathbb{R}^n, g)$, while it is clearly satisfiable in some DTM. According to Kremer and Mints (2005), the term

$$\mathbf{I}p \;\to\; \mathbf{C}\Diamond_F\mathbf{I}p,$$

where $\tau_1 \to \tau_2 = \overline{\tau_1} \sqcup \tau_2$, is valid in all unit balls, but refuted in a DTM based on an Aleksandrov space and a DTM based on $\mathbb{R}^n$ with the homeomorphism $(x_1, \dots, x_n) \mapsto (x_1, \dots, x_{n-1}, x_n + 1)$. Finally, the $\mathcal{DTL}$-term

$$\Box_F\mathbf{I}p \;\to\; \mathbf{I}\Box_F p$$

is valid in DTMs based on Aleksandrov spaces, but refuted in the classes of DTMs based on Euclidean spaces and unit balls.

DTMs with continuous functions. Theorem 9.38 shows that DTMs with *homeomorphisms* behave similarly to topological snapshot models. This situation changes drastically for DTMs with *continuous functions* (which are not necessarily open). In this case, no corresponding snapshot tt-models have been developed. To clarify—at least to some extent—the relation between the two kinds of models, let us consider DTMs based on Aleksandrov spaces.

Suppose that an Aleksandrov topological space $\mathfrak{T}_{\mathfrak{G}} = (W, \mathbb{I}_{\mathfrak{G}})$ is induced by the quasi-order $\mathfrak{G} = (W, R)$ (see Sec. 3.1). Then it is easy to check that a function $g\colon W \to W$ is a *continuous function* on $\mathfrak{T}_{\mathfrak{G}}$ iff for all $u, v \in W$,

$$uRv \qquad \text{implies} \qquad g(u)Rg(v).$$

(A bijection f is a *homeomorphism* on $\mathfrak{T}_{\mathfrak{G}}$ iff both the above implication and its converse hold.)

This observation suggests that DTMs based on Aleksandrov spaces with continuous functions correspond to what may be called *Aleksandrov snapshot models with expanding domains*. Indeed, suppose that $\mathfrak{A} = (\mathfrak{M}, g)$ is a DTM where $\mathfrak{M} = (\mathfrak{T}_{\mathfrak{G}}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots)$, $\mathfrak{G} = (W, R)$ is as above and g is a continuous and surjective map on $\mathfrak{T}_{\mathfrak{G}}$. Consider the sequence of models

$$\mathfrak{M}_0 = \mathfrak{M}, \quad \mathfrak{M}_1 = (\mathfrak{T}_{\mathfrak{G}_1}, p_0^{\mathfrak{M}_1}, \dots), \quad \mathfrak{M}_2 = (\mathfrak{T}_{\mathfrak{G}_2}, p_0^{\mathfrak{M}_2}, \dots), \dots \tag{9.31}$$

where

- $\mathfrak{G}_n = (W, R_n)$,
- uR_nv iff $g^n(u)Rg^n(v)$ for any $u, v \in W$,
- $u \in p_i^{\mathfrak{M}_n}$ iff $g^n(u) \in p_i^{\mathfrak{M}}$.

The temporal and topological operators on this sequence of models can be interpreted in exactly the same way as in Sec. 6.1. In particular,

- $u \in (\mathbf{C}\tau)^{\mathfrak{M}_n}$ iff there is $v \in W$ such that uR_nv and $v \in \tau^{\mathfrak{M}_n}$,
- $u \in (\Diamond_F\tau)^{\mathfrak{M}_n}$ iff there is $m > n$ such that $u \in \tau^{\mathfrak{M}_m}$.

We then obtain that, for every $\mathcal{DTL}$-term τ, every $w \in W$ and every $n \geq 0$,

$$g^n(w) \in \tau^{\mathfrak{A}} \quad \text{iff} \quad w \in \tau^{\mathfrak{M}_n},$$

and so τ is satisfiable in $\mathfrak{A}$ iff τ is satisfiable in (9.31).

The difference between (9.31) and the snapshot models we have considered before is that the spaces $\mathfrak{T}_{\mathfrak{G}_n}$ or, which is the same, the quasi-orders $\mathfrak{G}_n = (W, R_n)$ do not necessarily coincide. More precisely, using the fact that g is continuous it is easy to see that $R_n \subseteq R_{n+1}$ for every $n \geq 0$; see Fig. 9.10. That is why we call these models *snapshot models with expanding domains*. Fig. 9.10 also shows that the term

$$\bigcirc\mathbf{C}\tau \to \mathbf{C}\bigcirc\tau$$

is not valid in all DTMs with continuous functions, while it is clearly valid in all DTMs with homeomorphisms. For more details on the connection between such models and DTMs based on Aleksandrov spaces with continuous functions see Gabelaia et al., 2006.

It is known (e.g., Gabbay et al., 2003) that satisfiability in models with expanding domains can be reduced to satisfiability in models with constant domains, but not the other way round as we shall see a bit later. So in principle one could expect that the dynamic topological logics interpreted in DTMs based on

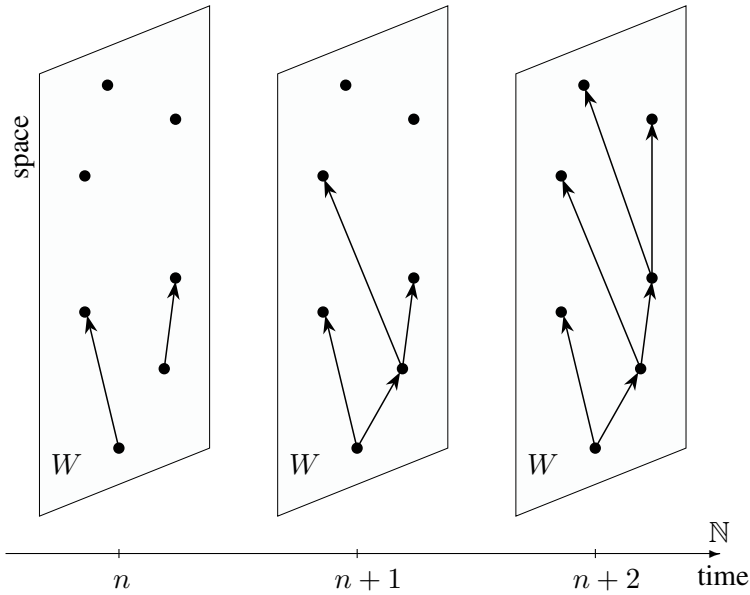


Figure 9.10. Model with expanding domains.

arbitrary, Aleksandrov or Euclidean topological spaces with continuous functions behave 'better' than their counterparts with homeomorphisms. Indeed, a fine-grained complexity analysis reveals interesting differences between the logic of homeomorphisms and the logic of continuous functions. We begin with the following 'negative' theorem proved in Konev et al., 2005 and Konev et al., 2006a:

THEOREM 9.40 *Let $\mathcal{M}$ be any of the following classes of dynamic topological models:*

- *DTMs based on Aleksandrov spaces with continuous functions;*
- *DTMs based on topological spaces with continuous functions;*
- *DTMs based on $\mathbb{R}^n$ with continuous functions, for $n \geq 1$.*

Then the satisfiability problem for $\mathcal{DTL}$-terms in $\mathcal{M}$ is undecidable.

Note that, in contrast to DTMs with homeomorphisms, it is still not clear whether any of these logics is recursively enumerable or even finitely axiomatisable. However, the first exciting difference between the algorithmic behaviour of the two models can be observed by considering the fragment of $\mathcal{DTL}$ in

which the topological operators are not applied to formulas containing the 'infinitary' temporal operators $\Box_F$ and $\Diamond_F$. This language is still very expressive and the undecidability/non-axiomatisability results of Theorems 9.39 and 9.40 still hold for it. However, the set of formulas from this fragment that are valid in DTMs based on Aleksandrov spaces or arbitrary topological spaces with continuous functions is recursively enumerable. This is proved in Konev et al., 2006a by an application of Kruskal's tree theorem.

The proof of Theorem 9.40 proceeds by a rather involved reduction of the ω-reachability problem for lossy channel systems (Schnoebelen, 2002). It essentially uses the fact that the number of function iterations is infinite. This observation opens a second possibility for a fine-grained complexity analysis: what happens if we consider DTMs where only *finitely* (but unboundedly) many function iterations are allowed. In this case the interpretation of $\mathcal{DTL}$-terms containing temporal operations depends of course on the iteration step of g.

More precisely, let $\mathfrak{A} = (\mathfrak{M}, g)$ be a DTM based on a topological space $(U, \mathbb{I})$, $N > 0$ is the allowed number of iterations of g, and $n \leq N$. Given a $\mathcal{DTL}$-term τ, we define $\tau^{\mathfrak{A},n,N}$, the *extension of τ after n steps in the DTM $\mathfrak{A}$ with N iterations*, inductively as follows:

- $p_i^{\mathfrak{A},n,N} = p_i^{\mathfrak{M}}$,
- $(\tau_1 \sqcap \tau_2)^{\mathfrak{A},n,N} = \tau_1^{\mathfrak{A},n,N} \cap \tau_2^{\mathfrak{A},n,N}$,
- $(\overline{\tau})^{\mathfrak{A},n,N} = U - \tau^{\mathfrak{A},n,N}$,
- $(\mathbf{I}\tau)^{\mathfrak{A},n,N} = \mathbb{I}\tau^{\mathfrak{A},n,N}$,
- $(\bigcirc\tau)^{\mathfrak{A},n,N} = \emptyset$ for $n = N$, and $(\bigcirc\tau)^{\mathfrak{A},n,N} = g^{-1}(\tau^{\mathfrak{A},n+1,N})$ otherwise,
- $(\Diamond_F\tau)^{\mathfrak{A},n,N} = \bigcup_{m=n+1}^{N} g^{n-m}(\tau^{\mathfrak{A},m,N})$.

Say that τ is *satisfiable in DTMs from a class $\mathcal{M}$ with finite iterations*, or *fi-satisfiable in $\mathcal{M}$*, for short, if there exist a DTM $\mathfrak{A}$ in $\mathcal{M}$ and $N > 0$ such that $\tau^{\mathfrak{A},0,N} \neq \emptyset$.

It is not hard to see that the reduction of Post's correspondence problem from the proof of Theorem 9.39 can be also used to prove the following:

THEOREM 9.41 *Let $\mathcal{M}$ be any of the classes of DTMs mentioned in Theorem 9.38. Then fi-satisfiability of $\mathcal{DTL}$-terms in $\mathcal{M}$ is undecidable.*

On the contrary, if we consider the class of DTMs based on arbitrary topological spaces with continuous functions then one can first reduce fi-satisfiability in this class to fi-satisfiability in DTMs based on finite Aleksandrov spaces

with continuous functions, and then use Kruskal's tree theorem to prove the following:

THEOREM 9.42 *Let $\mathcal{M}$ be one of the following classes:*

- *DTMs based on Aleksandrov spaces with continuous functions,*
- *DTMs based on topological spaces with continuous functions.*

Then fi-satisfiability of $\mathcal{DTL}$-terms in $\mathcal{M}$ is decidable, but not in primitive recursive time.

The non-primitive recursive lower bound is proved by reduction of the reachability problem for lossy channel systems. All details can be found in Gabelaia et al., 2006.

8.2 Dynamic metric logics

In this section we consider our fourth formal model—dynamic metric systems. Similarly to dynamic topological models from Sec. 8.1, a *dynamic metric model* (DMM, for short) is a pair of the form

$$\mathfrak{A} = (\mathfrak{M}, g),$$

where $\mathfrak{M} = (\mathfrak{D}, p_0^{\mathfrak{M}}, p_1^{\mathfrak{M}}, \dots)$ is a metric model based on a metric space $\mathfrak{D} = (\Delta, d)$ and g is a function on $\mathfrak{D}$. We will only consider *isometric functions*, i.e., bijections on Δ such that $d(x, y) = d(g(x), g(y))$, for all $x, y \in \Delta$. For instance, the translation $x \mapsto x + 1$ and reflection $x \mapsto -x$ maps on $\mathbb{R}$, the rotations g_α of the two-dimensional unit ball $B^2 = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 \leq 1\}$ by the angle α around $(0, 0)$ are isometric automorphisms on the respective spaces.

We only consider the simplest language $\mathcal{DML}^{\leq}$ of *dynamic metric logic*, which is defined in the same way as $\mathcal{DTL}$ with the exception that the topological operators are now replaced by the *metric operators* $\exists^{\leq a}$ and $\forall^{\leq a}$, for $a \in \mathbb{Q}^{\geq 0}$. Formally, $\mathcal{DML}^{\leq}$-terms are

$$\tau \quad ::= \quad p \quad | \quad \overline{\tau} \quad | \quad \tau_1 \sqcap \tau_2 \quad | \quad \exists^{\leq a}\tau \quad | \quad \bigcirc\tau \quad | \quad \Box_F\tau \quad | \quad \Diamond_F\tau.$$

Again, we omit the 'until' operator to keep our language simple, although all the results can be extended to the language including 'until.'

In a dynamic metric model $\mathfrak{A} = (\mathfrak{M}, g)$, terms τ are interpreted as sets $\tau^{\mathfrak{A}} \subseteq \Delta$, where $\mathfrak{M}$ is based on the metric space $\mathfrak{D} = (\Delta, d)$. For spatial variables we have $p_i^{\mathfrak{A}} = p_i^{\mathfrak{M}}$; the Boolean operators are interpreted as usual, the metric operators $\exists^{\leq a}\tau$ and $\forall^{\leq a}\tau$ as in Sec. 3.3, and the temporal operators as in Sec. 8.1.

The notions of satisfiability and validity of $\mathcal{DML}^{\leq}$-terms are defined in the standard way. The next theorem connects satisfiability in DMMs with satisfiability in snapshot models from Sec. 7:

THEOREM 9.43 *A $\mathcal{DML}^{\leq}$-term τ is satisfiable in a DMM with an isometric function iff the formula $\neg(\tau = \bot)$ is satisfiable in a metric snapshot model based on the same metric space.*

As it happened with metric temporal logics in Sec. 7, dynamic metric logics are slightly simpler then their topological counterparts:

THEOREM 9.44 *The set of $\mathcal{DML}^{\leq}$-terms that are valid in DMMs with isometric functions is decidable. However, the decision problem is not elementary.*

This theorem should not come as a surprise: its claim and the proof are essentially the same as those of Theorem 9.34 (all details can be found in Konev et al., 2006b).

9. Related 'temporalised' formalisms

The logics we have considered in this chapter can be regarded as *temporalisations* of static spatial logics. As many other 'static' logics have also been extended by a temporal dimension, for example, first-order temporal logic (Gabbay et al., 1994), temporal epistemic logic (Fagin et al., 1995), temporal description logic (Gabbay et al., 2003), it makes sense to briefly discuss similarities and differences between these temporalisations.

The most generic approach to the temporalisation of a static logic is of course *first-order temporal logic*. In this logic, temporal operators may occur anywhere in first-order formulas (in particular, in the scope of quantifiers), and the intended models are flows of time where each time point is represented by a relational structure interpreting the first-order part of the language. It is known since the 1960s that the resulting logics are extremely complex, mostly Σ^1_1-complete (see, e.g., Gabbay et al., 1994, Gabbay et al., 2003 and references therein). For example, the two-variable fragment, the monadic fragment, and the guarded fragment of first-order temporal logic over the natural numbers and with constant or expanding domains is Σ^1_1-complete.

Only recently the so-called *monodic* fragments of first-order temporal logics (in which temporal operators are only applied to formulas with at most one free variable) have been identified as expressive yet often decidable (or at least recursively enumerable) fragments (Hodkinson et al., 2000; Hodkinson et al., 2001; Gabbay et al., 2003). The positive results about the monodic fragments rely, however, on the fact that they are not able to express that a *binary* relation does not change over time. In other words, in the monodic fragments one can reason about the change (or non-change) of unary predicates but not about the change (or non-change) of binary relations. This feature of monodic fragments

is in sharp contrast with the logics we encounter in the context of spatio-temporal representation and reasoning: as we have seen, in this case we usually expect the underlying space (e.g., a metric or topological space) not to change in time. What changes is the extension of unary predicates. That is to say, we almost always have at least one constant binary relation (or higher-order operator): in metric spaces the relation $R(x, y)$ defined by $d(x, y) < a$, in Aleksandrov spaces the relation R inducing the topological space, in arbitrary topological spaces even the higher-order interior operator, etc. For this reason, *the results on the decidability of monodic fragments do not apply to spatio-temporal logics*. In fact, we have seen that the straightforward combination of spatial and temporal formalisms almost always leads to highly undecidable logics. In the more abstract setting of *products of modal logics* this phenomenon has been recently investigated by Gabbay et al. (2003) and Gabelaia et al. (2005b, 2006).

The main message to be deduced from the results on combinations of spatial and temporal formalisms is that a fine-tuned analysis of both the spatial logic and the interaction between spatial and temporal operators is required in order to obtain expressive and still decidable formalisms. There appears to be no general way of translating positive results from other temporalisations to spatio-temporal logics. Actually, this is also the case for temporal epistemic logic and temporal description logic. Again, most of the 'positive' results in those areas depend on the assumption that one cannot reason about the change (and non-change) of binary relations. With one exception, the results in those areas are therefore much closer to the results on monodic fragments of first-order temporal logics than to the results on spatio-temporal logics.

The only exception from this rule we know is the decidability (in non-elementary time) of the satisfiability problem for terms of the metric temporal and dynamic logics. Although this result cannot be obtained as an instance of a known result from other temporalised formalisms, its proof nevertheless closely resembles the proofs of the following results:

- The decidability (in non-elementary time) of the temporal epistemic logic with multi-modal $\mathcal{S}5$ interpreted in synchronous systems with perfect recall and no learning (Halpern and Vardi, 1989).
- In temporal description logic, the decidability (in non-elementary time) of the satisfiability problem for temporalised $\mathcal{ALC}$ where roles (binary predicates) do not change over time (Wolter and Zakharyaschev, 1999; Gabbay et al., 2003).

In all these cases, we deal with models where certain relations do not change over time (in the epistemic case these are the equivalence relations interpreting the epistemic operators, in the description logic case these are the roles interpreting the value restrictions). The crucial property underlying the decidability proofs is that those constant relations can be assumed to form tree-like structures and

that the satisfaction relation is 'local' in the sense that the interpretation of terms (propositional variables/concepts) in a certain distance from the root of the tree-like structure does not influence the satisfaction relation in the root.

Notice, however, that in each case one has to consider carefully the constraints on the relations. As we know from Theorem 9.21, a decidability proof does not go through for transitive relations (from Aleksandrov spaces) which do not change over time.

Acknowledgements

The work on this chapter was partially supported by the U.K. EPSRC grants GR/S61966, GR/S63182, GR/S63175, GR/S61973.

References

Aiello, M. and van Benthem, J. (2002). A modal walk through space. *Journal of Applied Non-Classical Logics*, 12(3–4):319–364.

Alexandroff, P. S. (1937). Diskrete Räume. *Matematicheskii Sbornik*, 2 (44): 501–518.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843.

Allen, J. (1984). Towards a general theory of action and time. *Artificial Intelligence*, 26:123–154.

Allouche, J.-P., Courbage, M., and Skordev, G. (2001). Notes on cellular automata. *Cubo, Matemática Educacional*, 3:213–244.

Areces, C., Blackburn, P., and Marx, M. (2000). The computational complexity of hybrid temporal logics. *Logic Journal of the IGPL*, 8:653–679.

Arhangel'skii, A. and Collins, P. (1995). On submaximal spaces. *Topology and its Applications*, 64:219–241.

Asher, N. and Vieu, L. (1995). Toward a geometry of common sense: A semantics and a complete axiomatization of mereotopology. In Mellish, C., editor, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 846–852. Morgan Kaufmann.

Balbiani, P. and Condotta, J.-F. (2002). Computational complexity of propositional linear temporal logics based on qualitative spatial or temporal reasoning. In Armando, A., editor, *Proceedings of Frontiers of Combining Systems (FroCoS 2002)*, volume 2309 of *Lecture Notes in Computer Science*, pages 162–176. Springer.

Balibiani, P., Tinchev, T., and Vakarelov, D. (2004). Modal logics for region-based theories of space. Manuscript.

Barwise, J., editor (1977). *Handbook of Mathematical Logic*. North-Holland, Amsterdam.

Bennett, B. (1994). Spatial reasoning with propositional logic. In *Proceedings of the 4th International Conference on Knowledge Representation and Reasoning*, pages 51–62. Morgan Kaufmann.

Bennett, B., Cohn, A., Wolter, F., and Zakharyschev, M. (2002). Multi-dimensional modal logic as a framework for spatio-temporal reasoning. *Applied Intelligence*, 17:239–251.

Blackburn, P. (1992). Fine grained theories of time. In Aurnague, M., Borillo, A., Borillo, M., and Bras, M., editors, *Proceedings of the 4th European Workshop on Semantics of Time, Space, and Movement and Spatio-Temporal Reasoning*, pages 299–320, Château de Bonas, France. Groupe 'Langue, Raisonnement, Calcul', Toulouse.

Bourbaki, N. (1966). *General Topology, Part 1*. Hermann, Paris and Addison-Wesley.

Brown, J. R. (1976). *Ergodic Theory and Topological Dynamics*. Academic Press.

Burgess, J. (1979). Logic and time. *Journal of Symbolic Logic*, 44:566–582.

Clarke, E. and Emerson, E. (1981). Design and synthesis of synchronisation skeletons using branching time temporal logic. In Kozen, D., editor, *Logic of Programs*, volume 131 of *Lecture Notes in Computer Science*, pages 52–71. Springer.

Clarke, E., Grumberg, O., and Peled, D. (2000). *Model Checking*. MIT Press.

Cohn, A. (1997). Qualitative spatial representation and reasoning techniques. In Brewka, G., Habel, C., and Nebel, B., editors, *KI-97: Advances in Artificial Intelligence*, volume 1303 of *Lecture Notes in Computer Science*, pages 1–30. Springer.

Cohn, A. G. and Hazarika, S. M. (2001). Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46:1–29.

Davis, E. (1990). *Representations of Commonsense Knowledge*. Morgan Kaufmann.

Egenhofer, M. and Franzosa, R. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5:161–174.

Egenhofer, M. J. and Herring, J. R. (1991). Categorizing topological relationships between regions, lines and point in geographic databases. Technical report, University of Maine.

Emerson, E. and Halpern, J. (1986). 'Sometimes' and 'not never' revisited: on branching versus linear time. *Journal of the ACM*, 33:151–178.

Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. MIT Press.

Finger, M. and Gabbay, D. (1992). Adding a temporal dimension to a logic system. *Journal of Logic, Language and Information*, 2:203–233.

Fisher, M., Gabbay, D., and Vila, L., editors (2005). *Handbook of Temporal Reasoning in Artificial Intelligence*. Elsevier.

Gabbay, D., Hodkinson, I., and Reynolds, M. (1994). *Temporal Logic: Mathematical Foundations and Computational Aspects, Volume 1*. Oxford University Press.

Gabbay, D., Kurucz, A., Wolter, F., and Zakharyaschev, M. (2003). *Many-Dimensional Modal Logics: Theory and Applications*, volume 148 of *Studies in Logic*. Elsevier.

Gabbay, D., Reynolds, M., and Finger, M. (2000). *Temporal Logic: Mathematical Foundations and Computational Aspects, Volume 2*. Oxford University Press.

Gabelaia, D., Kontchakov, R., Kurucz, A., Wolter, F., and Zakharyaschev, M. (2005a). Combining spatial and temporal logics: expressiveness vs. complexity. *Journal of Artificial Intelligence Research (JAIR)*, 23:167–243.

Gabelaia, D., Kurucz, A., Wolter, F., and Zakharyaschev, M. (2005b). Products of 'transitive' modal logics. *Journal of Symbolic Logic*, 70(3):993–1021.

Gabelaia, D., Kurucz, A., Wolter, F., and Zakharyaschev, M. (2006). Non-primitive recursive decidability of products of modal logics with expanding domains. *Annals of Pure and Applied Logic*, 142(1–3):245–268.

Gerevini, A. and Nebel, B. (2002). Qualitative spatio-temporal reasoning with RCC-8 and Allen's interval calculus: Computational complexity. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'02)*, pages 312–316. IOS Press.

Gödel, K. (1933). Eine Interpretation des intuitionistischen Aussagenkalküls. *Ergebnisse eines mathematischen Kolloquiums*, 4:39–40.

Goranko, V., Montanari, A., and Sciavicco, G. (2004). A road map of interval temporal logics and duration calculi. *Journal of Applied Non-Classical Logics*, 14:9–54.

Goranko, V. and Passy, S. (1992). Using the universal modality: gains and questions. *Journal of Logic and Computation*, 2:5–30.

Gotts, N. (1996). An axiomatic approach to topology for spatial information systems. Technical Report 96.25, School of Computer Studies, University of Leeds.

Halpern, J. and Vardi, M. (1989). The complexity of reasoning about knowledge and time I: lower bounds. *Journal of Computer and System Sciences*, 38: 195–237.

Hirshfeld, Y. and Rabinovich, A. (1999). Quantitative temporal logic. In *Proceedings of Computer Science Logic 1999*, pages 172–187. Springer.

Hodkinson, I., Wolter, F., and Zakharyaschev, M. (2000). Decidable fragments of first-order temporal logics. *Annals of Pure and Applied Logic*, 106:85–134.

Hodkinson, I., Wolter, F., and Zakharyaschev, M. (2001). Monodic fragments of first-order temporal logics: 2000–2001 A.D. In Nieuwenhuis, R. and Voronkov, A., editors, *Logic for Programming, Artificial Intelligence and*

Reasoning, volume 2250 of *Lecture Notes in Artificial Intelligence*, pages 1–23. Springer.

Hughes, G.E. and Cresswell, M.J. (1996). *A New Introduction to Modal Logic*. Methuen, London.

Kamp, H. (1968). *Tense Logic and the Theory of Linear Order*. PhD thesis, University of California, Los Angeles.

Katok, A. and Hasselblatt, B. (1995). *Introduction to Modern Theory of Dynamical Systems*, volume 54 of *Encyclopedia of mathematics and its applications*. Elsevier.

Konev, B., Kontchakov, R., Wolter, F., and Zakharyaschev, M. (2006a). Dynamic topological logics over spaces with continuous functions. In Hodkinson, I. and Venema, Y., editors, *Advances in Modal Logic, vol. 6*, pages 299–318. College Publications, London.

Konev, B., Kontchakov, R., Wolter, F., and Zakharyaschev, M. (2006b). On dynamic topological and metric logics. *Studia Logica*, 84:127–158.

Konev, B., Wolter, F., and Zakharyaschev, M. (2005). Temporal logics over transitive states. In Nieuwenhuis, R., editor, *Proceedings of CADE-05*, volume 3632 of *LNCS*, pages 182–203. Springer.

Kremer, P. and Mints, G. (2005). Dynamic topological logic. *Annals of Pure and Applied Logic*, 131:133–158.

Kutz, O., Sturm, H., Suzuki, N.-Y., Wolter, F., and Zakharyaschev, M. (2003). Logics of metric spaces. *ACM Transactions on Computational Logic*, 4: 260–294.

Lamport, L. (1980). 'Sometimes' is sometimes 'not never'. In *Proceedings of the 7th ACM Symposium on Principles of Programming Languages*, pages 174–185.

Lewis, C. and Langford, C. (1932). *Symbolic Logic*. Appleton-Century-Crofts, New York.

Manna, Z. and Pnueli, A. (1992). *The Temporal Logic of Reactive and Concurrent Systems*. Springer.

Manna, Z. and Pnueli, A. (1995). *Temporal Verification of Reactive Systems: Safety*. Springer.

McKinsey, J.C.C. (1941). A solution of the decision problem for the Lewis systems **S2** and **S4**, with an application to topology. *Journal of Symbolic Logic*, 6:117–134.

McKinsey, J.C.C. and Tarski, A. (1944). The algebra of topology. *Annals of Mathematics*, 45:141–191.

Muller, P. (1998). A qualitative theory of motion based on spatio-temporal primitives. In Cohn, A., Schubert, L., and Shapiro, S., editors, *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, pages 131–142. Morgan Kaufmann.

Nutt, W. (1999). On the translation of qualitative spatial reasoning problems into modal logics. In Burgard, W., Christaller, T., and Cremers, A., editors, *Advances in Artificial Intelligence. Proceedings of the 23rd Annual German Conference on Artificial Intelligence (KI'99)*, volume 1701 of *Lecture Notes in Computer Science*, pages 113–124. Springer.

Ono, H. and Nakamura, A. (1980). On the size of refutation Kripke models for some linear modal and tense logics. *Studia Logica*, 39:325–333.

Orlov, I. (1928). The calculus of compatibility of propositions. *Mathematics of the USSR, Sbornik*, 35:263–286. (In Russian).

Pratt-Hartmann, I. (2002). A topological constraint language with component counting. *Journal of Applied Non-Classical Logics*, 12(3–4):441–467.

Prior, A. (1968). Now. *Noûs*, 2:101–119.

Randell, D., Cui, Z., and Cohn, A. (1992). A spatial logic based on regions and connection. In Nebel, B., Rich, C., and Swartout, W., editors, *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, pages 165–176. Morgan Kaufmann.

Renz, J. (1998). A canonical model of the region connection calculus. In Cohn, A., Schubert, L., and Shapiro, S., editors, *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning (KR'98)*, pages 330–341. Morgan Kaufmann.

Renz, J. and Nebel, B. (1999). On the complexity of qualitative spatial reasoning. *Artificial Intelligence*, 108:69–123.

Reynolds, M. (2002). Axioms for branching time. *Journal of Logic and Computation*, 12(4):679–697.

Schnoebelen, Ph. (2002). Verifying lossy channel systems has nonprimitive recursive complexity. *Information Processing Letters*, 83:251–261.

Sheremet, M., Tishkovski, D., Wolter, F., and Zakharyaschev, M. (2006). From topology to metric: modal logic and quantification in metric spaces. In Hodkinson, I. and Venema, Y., editors, *Advances in Modal Logic, vol. 6*, pages 429–448. College Publications, London.

Sheremet, M., Tishkovsky, D., Wolter, F., and Zakharyaschev, M. (2005a). 'Closer' representation and reasoning. In Horrocks, I., Sattler, U., and Wolter, F., editors, *International Workshop on Description Logics, (DL 2005)*, pages 25–36.

Sheremet, M., Tishkovsky, D., Wolter, F., and Zakharyaschev, M. (2005b). Comparative similarity, tree automata, and Diophantine equations. In *Proceedings of LPAR 2005, volume 3835 of LNAI*, pages 651–665. Springer.

Sistla, A. and Clarke, E. (1985). The complexity of propositional linear temporal logics. *Journal of the Association for Computing Machinery*, 32:733–749.

Slavnov, S. (2003). Two counterexamples in the logic of dynamic topological systems. Technical Report TR–2003015, Cornell University.

Smith, T. and Park, K. (1992). An algebraic approach to spatial reasoning. *International Journal of Geographical Information Systems*, 6:177–192.

Stockmeyer, L. (1974). *The Complexity of Decision Problems in Automata Theory and Logic*. PhD thesis, MIT.

Stone, M. (1937). Application of the theory of Boolean rings to general topology. *Transactions of the AMS*, 41:321–364.

Tarski, A. (1938). Der Aussagenkalkül und die Topologie. *Fundamenta Mathematicae*, 31:103–134.

Thomason, R. (1984). Combinations of tense and modality. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 2, pages 135–165. Reidel, Dordrecht.

Tsao Chen, T. (1938). Algebraic postulates and a geometric interpretation of the Lewis calculus of strict implication. *Bulletin of the AMS*, 44:737–744.

Vilain, M., Kautz, H., and van Beek, P. (1989). Constraint propagation algorithms for temporal reasoning—a revised report. In Weld, D. S. and de Kleer, J., editors, *Readings in Qualitative Reasoning about Physical Systems*, pages 373–381. Morgan Kaufmann.

Wolter, F. and Zakharyaschev, M. (1999). Modal description logics: modalizing roles. *Fundamenta Informaticae*, 39:411–438.

Wolter, F. and Zakharyaschev, M. (2000). Spatial reasoning in RCC–8 with Boolean region terms. In Horn, W., editor, *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000)*, pages 244–248. IOS Press.

Wolter, F. and Zakharyaschev, M. (2002). Qualitative spatio-temporal representation and reasoning: a computational perspective. In Lakemeyer, G. and Nebel, B., editors, *Exploring Artifitial Intelligence in the New Millenium*, pages 175–216. Morgan Kaufmann.

Wolter, F. and Zakharyaschev, M. (2003). Reasoning about distances. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*, pages 1275–1280. Morgan Kaufmann.

Wolter, F. and Zakharyaschev, M. (2005a). A logic for metric and topology. *Journal of Symbolic Logic*, 70:795–828.

Wolter, F. and Zakharyaschev, M. (2005b). On the computational complexity of metric logics. Manuscript.

Zanardo, A. (1996). Branching-time logic with quantification over branches: the point of view of modal logic. *Journal of Symbolic Logic*, 61:1–39.

Chapter 10

DYNAMIC TOPOLOGICAL LOGIC

Philip Kremer
University of Toronto

Grigori Mints
Stanford University

Second Readers

Jen Davoren
University of Melbourne

Valentin Shehtman
Moscow State University & King's College, London

1. Introduction

Dynamic Topological Logic provides a context for studying the confluence of three research areas: the topological semantics for S4, topological dynamics, and temporal logic. In the topological semantics, a *model* is a topological space X together with a valuation function V assigning to each propositional variable a subset of X. Conjunction is interpreted as intersection, disjunction as union, and negation as complementation. If we interpret the necessity connective, $\Box$, as topological interior, the resulting modal logic is S4. It has axioms

$$\Box(A \supset B) \supset (\Box A \supset \Box B), \qquad \Box A \supset A, \qquad \Box A \supset \Box\Box A$$

with *modus ponens* and necessitation, $A/\Box A$, as inference rules. Thus S4 is a general logic of topological spaces. For a general and comprehensive discussion see for example Rasiowa and Sikorski, 1963.

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 565–606.
© 2007 *Springer.*

Topological dynamics studies the asymptotic properties of continuous maps on topological spaces (Walters, 1982, p. 118). Let a *dynamic topological system* be an ordered pair $\langle X, f\rangle$ where X is a topological space and f is a continuous function on X. We can think of the function f as moving the points in X in each discrete unit of time: x gets moved to fx and then to ffx and so on. In Dynamic Topological Logic, the system S4 is extended to a logic of dynamic topological systems, by adding temporal modalities suited to formalizing the action of f on X. In particular, we want to formalize both the transition from one discrete moment to next, as f acts, moment by moment, on the points in X; and the asymptotic behaviour of the function f.

We turn to ω–time temporal logic with two future-looking modalities: *next*, $\bigcirc$; and *henceforth*, $*$. Suppose that we ignore topological issues and represent discrete moments as natural numbers. Let an *interpretation* be an assignment of a truth value to each propositional variable at each moment. The Boolean connectives are given their standard interpretations. As for the modalities, the formula $\bigcirc A$ is true at the moment m iff A is true at the next moment $m+1$; and the formula $*A$ is true at the moment m iff A is true at the moment n, for each $n \geq m$. Note that $*A$ is thus equivalent to the infinite conjunction

$$A \,\&\, \bigcirc A \,\&\, \bigcirc^2 A \,\&\, \bigcirc^3 A \,\&\, \ldots$$

The Linear Time Temporal Logic LTL of $\bigcirc$ and $*$ (cf. van Benthem, 1995 for an introduction and history) can be axiomatized by the classical tautologies; S4 axioms for $*$; the following axioms,

$$\bigcirc(A \vee B) \equiv (\bigcirc A \vee \bigcirc B); \qquad \bigcirc\neg A \equiv \neg\bigcirc A;$$

$$\bigcirc *A \equiv *\bigcirc A; \qquad *A \supset \bigcirc A;$$

$$\text{The induction axiom: } (A \,\&\, *(A \supset \bigcirc A)) \supset *A;$$

and the rules of Modus Ponens, and necessitation for $*$ (cf. Goldblatt, 1992 for example).

In this chapter, we combine the topological modality and the two temporal modalities, to define trimodal logics of dynamic topological systems or *dynamical topological logics* (abbreviated DTLs). Let a *dynamic topological model* be an ordered triple $\langle X, f, V\rangle$, where $\langle X, f\rangle$ is a dynamic topological system and V is a valuation function assigning to each propositional variable a subset of X. If we think of the subsets of X as the *propositions*, then, as in the static topological semantics, $\Box P = Int(P)$, for propositions P, where Int is the topological interior operation. We interpret the temporal modalities $\bigcirc$ and $*$ using the function f. Suppose that, at moment m, the proposition P is true at the point fx, i.e. $fx \in P$. Then after f has acted on x once, P will be true at x. In other words, at the next moment $m+1$, the proposition P is true at the point x. So at moment m, the proposition $\bigcirc P$ is true at x, that is $x \in \bigcirc P$ iff

$fx \in P$ iff $x \in f^{-1}(P)$. Thus the *next* modality is interpreted by the preimage under f:

$$\bigcirc P = f^{-1}(P).$$

The interpretation of $*$ as an infinite conjunction leads to the definition:

$$*P = \cap_{n \geq 0} f^{-n}(P)$$

Our goal is to investigate interesting classes of dynamical systems in this propositional framework. The restriction to a propositional framework is made for three reasons. First, this is the most natural way to extend the extensive work done on propositional modal logics: work on their axiomatizability, expressive power, and so on. Second, propositional systems are much more manageable than extensions admitting quantifiers. Third, it is interesting to investigate the expressive power of a purely propositional language: for example the *Poincaré Recurrence Theorem* can be stated in this language—cf. Sec. 3. In fact, this is the principal example that started our investigation: finding a manageable DTL of measure-preserving transformations is one of the principal unsolved problems in dynamical topological logic. We try to avoid, as far as possible, generalizations so wide that they are unlikely, given our present state of knowledge, to lead to results useful in mainstream mathematics.

As shown in (Konev et al., 2006a), the presence of $*$, the henceforth connective, often leads to incompleteness for a straightforward semantics. In particular, a significant range of DTLs are not axiomatizable: the DTL of homeomorphisms, the DTL of homeomorphisms on $\mathbb{R}^n$ (for any fixed $n \geq 1$), the DTL of homeomorphisms on Alexandrov spaces (see below), and the DTL of measure-preserving homeomorphisms on the unit ball of dimension n, where $n \geq 2$. Konev et al. (2006a) leave open the axiomatizability problem of DTLs that are based on continuous functions in general, rather than homeomorphisms.

Given the results in Konev et al., 2006a, most of the positive results for DTL are for a fragment of the trimodal language. Two extreme cases are $\Box$-free purely temporal formulas, and purely modal formulas without temporal connectives. In Sec. 4, we show that most DTLs are conservative over each of these fragments.

We also have a number of completeness results for the $\bigcirc\Box$ fragments of the language. For these results, we define two axiomatic systems: the logic of continuous functions S4C, introduced in Artemov et al., 1997, and the logic of homeomorphisms S4$\bigcirc$, introduced in Kremer and Mints, 1997. S4C is defined by the following axioms and rules:

Axioms the classical tautologies,
S4 axioms for $\Box$,
$\bigcirc(A \vee B) \equiv (\bigcirc A \vee \bigcirc B)$,
$\bigcirc\neg A \equiv \neg\bigcirc A$,

the continuity axiom: $\bigcirc\Box A \supset \Box\bigcirc A$

Rules Modus Ponens: $A, (A \supset B)/B$
Necessitation for $\bigcirc$: $A/\bigcirc A$
Necessitation for $\Box$: $A/\Box A$.

S4$\bigcirc$is defined by the axioms and rules of S4C, plus the following:

the converse of the continuity axiom: $\Box\bigcirc A \supset \bigcirc\Box A$

We will use S4C and S4$\bigcirc$ both for these axiomatizations and for the sets of all formulas derivable from the axioms by the inference rules.

Our main completeness results for the $\bigcirc\Box$ fragment of the language are as follows:

Sec. 6: S4$\bigcirc$ is sound and complete for

- homeomorphisms on topological spaces
- homeomorphisms on Alexandrov spaces
- homeomorphisms on finite topological spaces
- homeomorphisms on Cantor Space (This result is noted in Sec. 7.)
- homeomorphisms on $\mathbb{R}$
- homeomorphisms on the open [closed] unit interval

Sec. 7: S4C is sound and complete for

- continuous functions on topological spaces
- continuous functions on Alexandrov spaces, i.e. monotonic functions on Kripke frames (see Definitions 10.2 and 10.10)
- continuous functions on finite topological spaces
- continuous functions on Cantor Space

Most of our completeness results for mathematically interesting spaces are based on the topological completeness of S4 in the real line $\mathbb{R}$—a classic result of McKinsey and Tarski, 1944. In Sec. 5, we give a simplified proof of this classic result, based on a specially simple proof for rational numbers in the interval $(0, 1)$; this is followed, for real numbers in $(0, 1)$, by a passage to limit. Our main use of this result in is Sec. 6: our proof that S4$\bigcirc$ is complete for homeomorphic transformations on $\mathbb{R}$ proceeds first by giving a simple reduction

of an arbitrary formula (not containing $*$) to a formula that is almost $\bigcirc$-free; this reduction allows us to appeal to the completeness of S4 in $\mathbb{R}$.

This result does not extend to the logic S4C (Sec. 7): S4C is not complete for arbitrary continuous function on $\mathbb{R}$, as shown by counterexamples due to P. Kremer (2004) and S. Slavnov (2003). However we *are* able to prove completeness for Cantor Space (Sec. 7.5), an important subspace of $\mathbb{R}$. There are also two positive results in the vicinity: Slavnov, 2005 contains a proof that S4C is complete for arbitrary continuous functions on any $\mathbb{R}^n$, $n \geq 1$. To be more precise, S4C is complete for the follow set of dynamic topological systems: $\{\langle \mathbb{R}^n, f\rangle : n \geq 1$ and f is a continuous function on $\mathbb{R}^n\}$ and Fernandez, 2006 contains a proof that S4C is complete for arbitrary continuous functions on $\mathbb{R}^2$. Whether the logic of continuous functions on $\mathbb{R}$ is axiomatizable remains open.

We also note the completeness of a special axiomatization of a *temporal over topological* fragment where the temporal modalities cannot occur in the scope of a topological modality. See Kremer and Mints, 1997 and more recent results on recursive axiomatizability of the topological over temporal fragment in Konev et al., 2006b. Ch. 9 in this Handbook contains a great deal of material on the interaction between spatial and temporal logic.

The current chapter is part of a research program whose first results were announced in three conference abstracts (Kremer, 1997, Kremer and Mints, 1997, and Kremer et al., 1997). An independent and closely related research program saw its first results published in Artemov et al., 1997, and has been further pursued in Davoren, 1998.

2. Basic definitions

We work with a trimodal language L with a set PV of propositional variables; Boolean connectives $\vee$ and $\neg$; and three one-place modalities $\Box$ (interior), $\bigcirc$ (next) and $*$ (henceforth). We assume that $\&$, $\supset$ and $\equiv$, are defined in terms of $\vee$ and $\neg$. We use p, q, r as metavariables over PV and A, B, C as metavariables over formulas. The language $L^{\Box}$ is the fragment of L whose only modality is $\Box$; and the language $L^{\bigcirc\Box}$ is the fragment of L whose only modalities are $\bigcirc$ and $\Box$.

DEFINITION 10.1 *A* topological model *is an ordered pair, $M = \langle X, V\rangle$, where X is a topological space and $V : PV \to \mathcal{P}(X)$ is a valuation function assigning a subset of X to each propositional variable. The valuation V is extended to all formulas of $L^{\Box}$ as follows:*

$$\begin{aligned} V(A \vee B) &= V(A) \cup V(B), \\ V(\neg B) &= X - V(B), \textit{and} \\ V(\Box B) &= Int(V(B)), \end{aligned}$$

where the interior of a subset $Y \subseteq X$ *is denoted by* $Int(Y)$. *A formula* A *is* validated by $M = \langle X, V \rangle$ *iff* $V(A) = X$. *Notation:* $M \models A$.

DEFINITION 10.2 *A* Kripke frame *is an order pair* $\langle W, R \rangle$ *where* W *is a non-empty set (of* worlds*) and* R *is a reflexive and transitive relation on* W. *World* w' *is a* successor *of world* w *iff* wRw'. w *is* R-equivalent *to* w' *(in symbols,* $w \equiv_R w'$*) iff* wRw' *and* $w'Rw$.

DEFINITION 10.3 *Given a Kripke frame* $\langle W, R \rangle$*, a subset* S *of* W *is* open *iff* S *is closed under* R*: for every* $x, y \in W$*, if* $x \in S$ *and* xRy *then* $y \in S$. *The family of open sets forms a topology. Thus, for every Kripke frame* $\langle W, R \rangle$*, a topological space is defined by imposing that topology on the set* W. *We define the interior of* $X \subseteq W$*:*

$$Int(X) =_{\mathrm{df}} \{w \in W : \forall\, w' \in W, \textit{if } wRw' \textit{ then } w' \in X\}.$$

Note that, in these spaces, the intersection of arbitrary open sets is open: thus they are *Alexandrov* spaces (cf. Alexandrov, 1937) as defined in Artemov et al., 1997. In fact, as noted in Artemov et al., 1997 every Alexandrov space is generated in a natural way by a Kripke frame, so we can identify Kripke frames and Alexandrov spaces.

The next result (McKinsey and Tarski, 1944,Kripke, 1963) provides completeness for purely modal formulas.

THEOREM 10.4 (MCKINSEY-TARSKI-KRIPKE) *Suppose that* X *is a dense-in-itself metric space and* A *is a formula that does not contain the temporal connectives* $\bigcirc$ *and* $*$. *Then the following are equivalent:*

(i) $A \in$ S4.

(ii) $\models A$.

(iii) $X \models A$.

(iv) $\mathbb{R} \models A$.

(v) $Y \models A$ *for every finite topological space* Y.

(vi) $Y \models A$ *for every Alexandrov space* Y.

Proof The equivalence of (i)–(v) is due to McKinsey and Tarski, 1944. For the completeness of S4 in the real line, see a streamlined proof in Sec. 5. The equivalence of (i) and (vi) is due, in effect, to Kripke, 1963. QED

Thus not only does the topological interpretation give a semantics for S4, but S4 is the topological logic of a host of particular topological spaces, for example the real line, $\mathbb{R}$; the closed unit interval, $[0, 1]$; and any other dense-in-itself

metric space. So the purely modal language is expressively weak, unable, for example, to distinguish between $\mathbb{R}$ and $[0, 1]$ despite their topological dissimilarities. Part of the DTL project is to see whether analogues of Theorem 10.4 can be proved or disproved: see Sec. 6 and Sec. 7 below.

DEFINITION 10.5 *A* dynamic topological system *(DTS, cf. Brown, 1976, Furstenberg, 1981) is an ordered pair, $\langle X, f\rangle$, where X is a topological space and f is a continuous function on X. A* dynamic topological model *(DTM) is an ordered triple $M = \langle X, f, V\rangle$ where $\langle X, f\rangle$ is a DTS and $V : PV \to \mathcal{P}(X)$ is a valuation function assigning a subset of X to each propositional variable. The valuation V is extended to all formulas by the clauses in Definition 10.1 plus the following:*

$$\begin{aligned} V(\bigcirc B) &= f^{-1}(V(B)), \textit{and} \\ V(*B) &= \cap_{n\geq 0} f^{-n}(V(B)). \end{aligned}$$

A formula A is validated by $M = \langle X, f, V\rangle$ *iff $V(A) = X$. Notation: $M \models A$.*

DEFINITION 10.6 *Suppose that $\langle X, f\rangle$ is a DTS. Validity relations are defined in a standard way.*

$$\begin{aligned} \langle X, f\rangle \models B \quad &\textit{iff} \quad M \models B \textit{ for every model } M = \langle X, f, V\rangle. \\ X \models B \quad &\textit{iff} \quad \langle X, f\rangle \models B \textit{ for every continuous function } f. \\ B \textit{ is } \text{valid } (\models B) \quad &\textit{iff} \quad X \models B \textit{ for every topological space } X. \end{aligned}$$

Our goal is to axiomatize interesting classes of topological spaces and continuous functions on these spaces.

DEFINITION 10.7 *Suppose that $\mathcal{F}$ is a class of functions so that each $f \in \mathcal{F}$ is a continuous function on some topological space. Suppose that $\mathcal{T}$ is a class of topological spaces.*

$$\begin{aligned} \mathcal{T}, \mathcal{F} \models B \quad &\textit{iff} \quad \textit{for every } f \in \mathcal{F} \textit{ and every } X \in \mathcal{T}, \textit{if } f \textit{ is a continuous} \\ &\qquad \textit{function on } X \textit{ then } \langle X, f\rangle \models B. \\ \mathcal{F} \models B \quad &\textit{iff} \quad \textit{for every topological space } X \textit{ and every } f \in \mathcal{F}, \textit{if } f \textit{ is a} \\ &\qquad \textit{continuous function on } X \textit{ then } \langle X, f\rangle \models B. \\ \mathcal{T} \models B \quad &\textit{iff} \quad X \models B \textit{ for every topological space } X \in \mathcal{T}. \end{aligned}$$

We assume that that in specifying a particular continuous function, we specify both the function itself as a set of ordered pairs, and the topological space on which we are taking it to act.

We are now ready to define various *Dynamic Topological Logics*, or DTLs.

DEFINITION 10.8 *For any class $\mathcal{T}$ of topological spaces and any class $\mathcal{F}$ of continuous functions, we define*

$$\begin{aligned} \mathrm{DTL}_{\mathcal{T},\mathcal{F}} &= \{A : \mathcal{T}, \mathcal{F} \models A\}. \\ \mathrm{DTL}_{\mathcal{T}} &= \{A : \mathcal{T} \models A\}. \\ \mathrm{DTL}_{\mathcal{F}} &= \{A : \mathcal{F} \models A\}. \end{aligned}$$

Most of our completeness results for interesting spaces and classes of spaces are based on completeness for finite spaces of a special kind.

DEFINITION 10.9 *A* dynamic Kripke frame *is an ordered triple $K = \langle W,\ R,\ f\rangle$ where $\langle W, R\rangle$ is a Kripke frame, and f is a function on W that is R-monotonic with respect to R, i.e.*

$$wRw' \Rightarrow fwRfw'.$$

The world $r \in W$ is a root world *of $\langle W,\ R,\ f\rangle$ iff both $fr = r$ and $\forall\ w \in W,\ rRw$. $\langle W,\ R,\ f\rangle$ is* rooted *iff there is some root world $r \in W$.*

Our dynamic Kripke frames are the continuous Kripke frames of (Artemov et al., 1997). They give rise to *dynamic Alexandrov systems* in an obvious way (cf Definition 10.3).

DEFINITION 10.10 *A* dynamic Kripke model *is a quartuple $M{=}\langle W,\ R,\ f,\ V\rangle$, where $\langle W,\ R,\ f\rangle$ is a dynamic Kripke frame and V is a valuation function assigning a subset of W to each propositional variable. The valuation V is extended to all formulas as follows:*

$$\begin{aligned} V(A \vee B) &= V(A) \cup V(B) \\ V(A \,\&\, B) &= V(A) \cap V(B) \\ V(\neg A) &= X - V(A) \\ V(\Box A) &= Int(V(A)) \\ V(\bigcirc A) &= f^{-1}(V(A)) \\ V(*A) &= \cap_{0 \leq n} \bigcirc^n V(A) \end{aligned}$$

$\langle W,\ R,\ f,\ V\rangle$ is rooted *iff $\langle W,\ R,\ f\rangle$ is rooted. A formula A is* validated by *$M = \langle W,\ R,\ f,\ V\rangle$ iff $V(A) = W$. Notation: $M \models A$.*

This is equivalent to the familiar definition for Kripke models:

$$w \in V(\Box A) \iff w' \models A \text{ for all } w' \text{ such that } wRw'$$

$$w \in V(\bigcirc A) \iff f(w) \in V(A)$$

Given a particular DTL, we will also be interested in its *fragments*.

DEFINITION 10.11 *If D is a dynamic topological logic, then the* purely topological *fragment of D is the $(\bigcirc, *)$-free fragment, that is, the set of formulas without $(\bigcirc, *)$ belonging to D. The* purely temporal *fragment of D is the $\Box$-free fragment. The*next-interior *fragment of D is the $*$-free fragment. We denote these logics as $D^{\Box}$ and $D^{\bigcirc *}$ and $D^{\bigcirc \Box}$, respectively.*

We are especially interested in three DTLs: the DTL of all dynamic topological systems, DTL_0; the DTL of homeomorphisms, $\mathrm{DTL}_{\mathcal{H}}$; and the DTL of measure-preserving functions on the closed unit circle. Less interesting mathematically $\mathrm{DTL}_{\mathcal{A}}$ of Alexandrov spaces is important as a starting point for investigation of other classes of DTL, since Alexandrov spaces are, in effect, Kripke frames. $\mathrm{DTL}_{\mathcal{H}}$ is nonaxiomatizable (Konev et al., 2006a), and the question of the axiomatizability of the other systems is still open.

Though the axiomatizability of $\mathrm{DTL}_{\mathcal{A}}$ remains an open question, it is easy to see that $\mathrm{DTL}_0 \subsetneq \mathrm{DTL}_{\mathcal{A}}$. For we have (10.1) and (10.2), below:

$$(*\Box p \supset \Box *p) \notin \mathrm{DTL}_0 \tag{10.1}$$

$$(*\Box p \supset \Box *p) \in \mathrm{DTL}_{\mathcal{A}}. \tag{10.2}$$

(10.2) follows from the fact that, in an Alexandrov space, the intersection of arbitrary open sets is open. To see (10.1), let $M = \langle \mathbb{R}, f, V \rangle$ where $f(x) = 2x$ and $V(p) = (-1, 1)$. Note that $V(\Box p) = (-1, 1)$ so that $f^{-n}(V(\Box p)) = (-1/2^n, 1/2^n)$. Thus $V(*\Box p) = \{0\}$. Similarly, $V(*p) = \{0\}$. So $V(\Box *p) = \emptyset$. So $M \not\models (*\Box p \supset \Box *p)$.

3. Recurrence and the DTL of measure-preserving continuous functions on the closed unit interval

A central motivation for this study is the phenomenon of recurrence in measure theory and topological dynamics, and the possibility of expressing this phenomenon in the framework of propositional logic. In fact, we can express recurrence in our trimodal language. Let us restrict attention to the segment $[0, 1]$ of reals which is already difficult enough (cf. Konev et al., 2006a).

Suppose that f is a function on a set X. Say that a point $x \in S$ is *recurrent* (for S) if $f^n(x) \in S$ for some $n > 1$. Let μ be the Lebesgue measure defined on subsets of the closed unit interval, $[0, 1]$. We say that S is *measurable* iff, for every set S', we have $\mu(S') = \mu(S \cap S') + \mu(S' - S)$. We say that a function f on $[0, 1]$ is *measure-preserving* iff $\mu(f^{-1}(S)) = \mu(S)$ for every measurable $S \subseteq [0, 1]$. Consider the following (non-essential) extension of the *Poincaré Recurrence Theorem* on [0,1] (see Walters, 1982):

THEOREM 10.12 *If f is a measure-preserving continuous function on $[0, 1]$ then the set of recurrent points of a non-empty open set $S \subseteq [0, 1]$ is dense in S.*

In order to express recurrence in our trimodal language, define the possibility connective $\Diamond$ as $\neg\Box\neg$, and the possibility connective # as $\neg*\neg$. These represent topological closure and "some time in the future", respectively. Consider the formula

$$(\Box p \supset \Diamond\bigcirc\#\Box p). \tag{10.3}$$

Let $\langle X, f\rangle$ be any dynamic topological system. Note that $\langle X, f\rangle \models (10.3)$ iff,

$$\forall \text{open}\, O \subseteq X : O \subseteq Cl\{x : \text{there is an}\, n \geq 1\, \text{such that}\, f^n x \in O\}. \tag{10.4}$$

By Theorem 10.12, (10.4) is true when $X = [0, 1]$ and f is any measure-preserving continuous function on $[0, 1]$. Thus, by Theorem 10.12, $\langle [0, 1], f\rangle \models (10.3)$ when f is any measure-preserving continuous function on $[0, 1]$. So, in some sense, (10.3) expresses the phenomenon of recurrence.

Thus the class $\mathcal{M}$ of measure-preserving functions on $[0, 1]$ is of interest. As we have just shown the formula (10.3) is in $DTL_{\mathcal{M}}$. The results of (Konev et al., 2006a) imply that the DTL of measure-preserving *homeomorphisms* on $[0, 1]$ is not axiomatizable: the axiomatizability of $DTL_{\mathcal{M}}$ remains open. One can try to approach these systems in the same way as first order arithmetic or second order logic: find a manageable axiomatizable fragment, strong enough to derive most mathematically interesting results that are statable in a given language.

3.1 A Simple Decidable DTL

To illustrate the existence of an interesting tractable DTL we prove that the logic $DTL_{\mathcal{HM}}$ of measure-preserving homeomorphisms of the real interval $[0, 1]$ is very easily reducible to S4 and hence decidable. The key observation is the next proposition.

LEMMA 10.13 *The only measures preserving homeomorphisms on* $[0, 1]$ *are the identity* $f(x) = x$ *and* $f(x) = 1 - x$.

Proof Since f is a homeomorphism, it takes all values in $[0, 1]$ and is strictly monotonic, say increasing for definiteness. Then $f(0) = 0$, since otherwise 0 is not a value of f. For every $x \in [0, 1]$ $\mu(f^{-1}[0, x]) = \mu([0, x]) = x$ and $f^{-1}[0, x] = [0, f^{-1}(x)]$. Hence $f^{-1}(x) = x$. Applying f to both parts, $f(x) = x$. QED

For every formula ϕ of S4C define two reducts corresponding to $f(x) = x$ and $f(x) = 1 - x$ repectively. Note that under $f(x) = x$ one has

$$\bigcirc A \iff *A \iff A$$

while $f(x) = 1 - x$ for all x implies $f(f(x)) = x$, hence

$$\bigcirc\bigcirc A \iff A, \text{ and } *A \iff A \,\&\, \bigcirc A \tag{10.5}$$

Define

- ϕ^0 := the result of erasing all occurrences of $\bigcirc$, $*$ and
- ϕ^1 := the results of replacing $\bigcirc$ and $*$ according to (10.5) including pushing $\bigcirc$ to atomic formulas and dropping all occurrences of $\bigcirc\bigcirc$ (cf. the operation g in Sec. 6).

The following theorem provides a decision algorithm for $DTL_{\mathcal{HM}}$. The notation S4 $\vdash \phi$ for a formula containing $\bigcirc$ means that $\bigcirc\psi$ is treated as a new atomic formula (cf. Sec. 6).

THEOREM 10.14 DTL$_{\mathcal{HM}}$ $\models \phi$ *iff* S4 $\vdash \phi^0$ & ϕ^1

Proof The direction from S4 to DTL$_{\mathcal{HM}}$ is obvious. In the opposite direction we have only to prove that

$$\text{S4} \nvdash \phi \text{ implies } \langle (0,1), f\rangle \not\models \phi \tag{10.6}$$

for $f(x) = 1 - x$ and every formula ϕ, where $\bigcirc$ is applied only to atoms—that is, where ϕ is constructed from propositional variables p and formulas $\bigcirc p$ by the Boolean connectives and $\Box$. We apply the same method as in the proof of the similar result for $\mathbb{R}$ and the logic of homeomorphisms (Theorem 10.46).

Find an S4-countermodel $M = \langle (\frac{1}{2}, 1), V\rangle$ for ϕ on the interval $(\frac{1}{2}, 1)$. Let $f(x) = 1 - x$, so that $f^{-1} = f$. Define a new valuation

$$V'(p) = V(p) \cup f^{-1}(V(\bigcirc p)) = V(p) \cup f(V(\bigcirc p))$$

on the dynamic topological system $\langle (0,1), f\rangle$. See Fig. 10.1. The same

Figure 10.1. $V'(p)$ is shaded.

computation as in the proof of Theorem 10.46 shows that

$$V(B) = I \cap V'(B) \text{ for all formulas B and } I = (\tfrac{1}{2}, 1).$$

The only thing to check anew is the implication

$$x \in I \text{ and } x \in V'(\bigcirc^n p) \Rightarrow x \in V(\bigcirc^n p) \text{ for } n = 0, 1.$$

We consider both cases, $n = 0$ and $n = 1$. Suppose that $n = 0$, and suppose that $x \in I$ and $x \in V'(\bigcirc^n p)$. Then $x \in V'(p)$. So either $x \in V(p)$ or $x \in f(V(\bigcirc p))$. But $I \cap f(V(\bigcirc p)) = \emptyset$. So $x \notin f(V(\bigcirc p))$. Thus $x \in V(p)$, as desired.

On the other hand, suppose that $n = 1$, and suppose that $x \in I$ and $x \in V'(\bigcirc^n p)$. Then $x \in V'(\bigcirc p)$. So $fx \in V'(p)$. So either $fx \in V(p)$ or $fx \in f(V(\bigcirc p))$. So either $x \in f^{-1}(V(p)) = f(V(p))$ or $x \in V(\bigcirc p)$. But $I \cap f(V(p)) = \emptyset$. So $x \notin f(V(p))$. Thus $x \in V(\bigcirc p)$, as desired. QED

4. Purely topological and purely temporal fragments of DTLs

In work on DTL, we foresee that most of the action will be in the interaction between the topological modality ($\Box$) and the temporal modalities ($\bigcirc$ and $*$). As it turns out, temporal differences often do not affect purely topological issues (see Theorem 10.15). Furthermore, the purely topological fragments and the purely temporal fragments of DTLs normally coincide with previously studied logics (see Theorems 10.16 and 10.20), namely S4 and LTL.

THEOREM 10.15 *Suppose that $\mathcal{T}$ is a class of topological spaces and $\mathcal{F}$ is a class of continuous functions. Also suppose that for every $X \in \mathcal{T}$, there is a $f \in \mathcal{F}$ with $dom(f) = X$. Then* $\mathrm{DTL}^{\Box}_{\mathcal{T},\mathcal{F}} = \mathrm{DTL}^{\Box}_{\mathcal{T}}$. *Thus temporal differences do not affect purely topological issues.*

Proof The inclusion $\supseteq$ is obvious. For $\subseteq$ take a modal formula $\alpha \notin \mathrm{DTL}^{\Box}_{\mathcal{T}}$ and a space $X \in \mathcal{T}$ with a continuos function g on X such that $\langle X, g\rangle \not\models \alpha$. Now replace g with a continuous function $f \in \mathcal{F}$ with the domain X. Since α is modal, $\langle X, f\rangle \models \alpha$ iff $\langle X, g\rangle \models \alpha$. QED

THEOREM 10.16 *Suppose that $\mathcal{T}$ is a class of topological spaces and that either*

(i) *every topological space is in $\mathcal{T}$,*

(ii) $\mathbb{R} \in \mathcal{T}$,

(iii) *some dense-in-itself metric space is in $\mathcal{T}$,*

(iv) *every finite topological space is in $\mathcal{T}$, or*

(v) *every Alexandrov space is in $\mathcal{T}$.*

Then $\mathrm{DTL}^{\Box}_{\mathcal{T}} = \mathrm{S4}$.

Proof This follows from the Theorem 10.15 and the McKinsey-Tarski-Kripke Theorem (Theorem 10.4). QED

COROLLARY 10.17

$$\mathrm{DTL}_0^{\Box} = \mathrm{DTL}_{\mathcal{H}}^{\Box} = \mathrm{DTL}_{\mathcal{M}}^{\Box} = \mathrm{DTL}_{\mathcal{A}}^{\Box} = \mathrm{DTL}_{\mathbb{R}}^{\Box} =$$

$$\mathrm{DTL}_{[0,1]}^{\Box} = \mathrm{DTL}_{\mathbb{R},\mathcal{H}}^{\Box} = \mathrm{DTL}_{\mathcal{A},\mathcal{H}}^{\Box} = \mathrm{DTL}_{fin}^{\Box} = \mathrm{S4},$$

where fin is the class of finite topological spaces.

DEFINITION 10.18 *Suppose that f is a continuous function and that $X = dom(f)$. For $m, n \in \omega$, f has the m–n–*property *iff there is some $x \in X$ such that $x, fx, ..., f^{m+n}x$ are all distinct and $f^{m+n+1}x = f^m x$. f has the ω–*property *iff there is some $x \in X$ such that $x, fx, f^2x, ...$ are all distinct. Suppose that $\mathcal{F}$ is a class of continuous functions. $\mathcal{F}$ is* rich *iff either (i) $\mathcal{F}$ contains some function with the ω–property or (ii) for each $m, n \in \omega$, $\mathcal{F}$ contains some function with the m–n–property.*

REMARK 10.19 *The following classes of functions are rich:*

(i) the class $\mathcal{H}$ of homeomorphisms;

(ii) the class $\mathcal{O}$ of open continuous functions (a function is *open* iff the image of every open set is open);

(iii) the class $\mathcal{M}$ of measure-preserving continuous functions on $[0, 1]$; and

(iv) the class of functions on finite topological spaces with the discrete topology.

For (i) and (ii) it suffices to find a homeomorphism on $\mathbb{R}$ with the ω–property, for example $fx = x + 1$. For (iii), the following function is continuous, measure-preserving and has the ω–property: $f(x) = 1 - 2x$ for $x \in [0, \frac{1}{2}]$ and $f(x) = 2x - 1$ for $x \in [\frac{1}{2}, 1]$. To see that f is measure-preserving consider any $S \subseteq [0, 1]$. Note that $\mu(f^{-1}(S) \cap [0, \frac{1}{2}]) = \mu(f^{-1}(S) \cap [\frac{1}{2}, 1]) = \frac{1}{2}\mu(S)$ so that $\mu(f^{-1}(S)) = \mu(S)$. See Fig. 10.2. To see that f has the ω–property, let $x = \sqrt{2} - 1$. Note that $f^n(x)$ is of the form $z \pm 2^n\sqrt{2}$, where z is an integer, so that $x, fx, f^2x, ...$ are all distinct. For (iv), we fix m and n and define a function with the m–n–property in the given class. Let X be the set $\{0, 1, 2, ..., m + n\}$ and let $fx = x + 1$ if $x < m + n$ and let $f(m + n) = m$.

THEOREM 10.20 *Suppose that $\mathcal{F}$ is a rich class of continuous functions. Then* $\mathrm{DTL}_{\mathcal{F}}^{\bigcirc *}$ = LTL.

Proof Recall the axiomatization of LTL given in the introduction. To show that LTL $\subseteq \mathrm{DTL}_{\mathcal{F}}^{\bigcirc *}$, it suffices to show that this axiomatization is sound for $\mathrm{DTL}_{\mathcal{F}}^{\bigcirc *}$. To show that $\mathrm{DTL}_{\mathcal{F}}^{\bigcirc *} \subseteq$ LTL, we consider two cases.

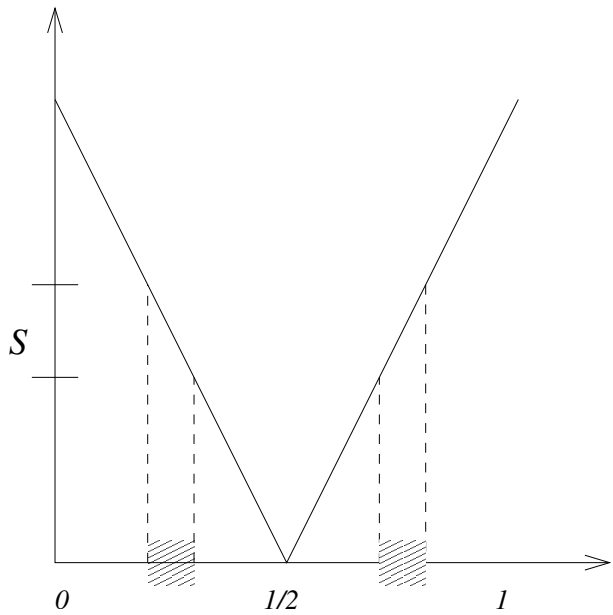


Figure 10.2. $f^{-1}(S)$ is shaded.

Case 1. $\mathcal{F}$ contains a function with the ω–property. Suppose that $A \notin$ LTL where A is in the language $L^{\bigcirc *}$. Then there is some infinite purely temporal model falsifying A. To be more precise, let an *infinite purely temporal model* be a function $V : PV \times \omega \to \{0, 1\}$, where PV is the set of propositional variables; where the natural numbers represent discrete moments in time; and where 0 and 1 represent falsity and truth. Given an infinite purely temporal model V, we define $n \models B$, for each $n \in \omega$ and each formula B in the language $L^{\bigcirc *}$ as follows: $n \models p$ iff $V(p, n) = 1$; $n \models \neg B$ iff $n \not\models B$; $n \models (B \vee C)$ iff $n \models B$ or $n \models C$; $n \models \bigcirc B$ iff $n+1 \models B$; and $n \models *B$ iff $m \models B$ for every $m \geq n$. The completeness theorem for LTL tells us that since $A \notin$ LTL , there is some infinite purely temporal model V such that $0 \not\models A$. Choose such a V.

Since $\mathcal{F}$ contains a function with the ω–property, we can choose a topological space X, a function $f \in \mathcal{F}$ and an $x \in X$, such that the points $x, fx, ffx, fffx, ...$ are all distinct. Choose a function $V' : PV \to X$ such that $f^k x \in V'(p)$ iff $V(p, k) = 1$, for every $k \in \omega$. Define $M = \langle X, T, V' \rangle$. By a standard induction on formulas, it can be shown that $f^k x \in V(B)$ iff $k \models B$ for all formulas B in the language $L^{\bigcirc *}$ and all $k \in \omega$. Thus $x \not\models A$ since $0 \not\models A$. So $A \notin \mathrm{DTL}_{\mathcal{F}}^{\bigcirc *}$, as desired.

Case 2. $\mathcal{F}$ contains a function with the m–n–property for every $m, n \in \omega$. Suppose that $A \notin$ LTL where A is in the language $L^{\bigcirc *}$. Let a *finite purely*

temporal model be an ordered triple $M = \langle Y, g, V\rangle$ where Y is a finite set; g is a function on Y; and $V : PV \times Y \to \{0, 1\}$. Given a finite purely temporal model $M = \langle Y, g, V\rangle$, we define $y \models B$, for each $y \in Y$ and each formula B in the language $L^{\bigcirc *}$ as follows: $y \models p$ iff $V(p, y) = 1$; $y \models \neg B$ iff $y \not\models B$; $y \models (B \vee C)$ iff $y \models B$ or $y \models C$; $y \models \bigcirc B$ iff $g(y) \models B$; and $y \models *B$ iff $g^n(y) \models B$ for every $n \geq 0$.

Segerberg, 1976 proves that LTL satisfies the finite frame property. So since $A \notin$ LTL , there is some finite purely temporal model $M = \langle Y, g, V\rangle$ and some $y \in Y$ such that $y \not\models A$. Since Y is finite,we have $g^{m+n+1}(y) = g^m(y)$, for some $m, n \in \omega$ with the $g^i(y)$ distinct for $i < m + n$. Choose such an m and n.

Choose a function $f \in \mathcal{F}$ with the m–n–property and let X be the topological space on which f acts. Choose an $x \in X$ such that $x, fx, ..., f^{m+n}x$ are all distinct, and such that $f^{m+n+1}x = f^m x$. Define $V' : PV \to \mathcal{P}(X)$ as follows:

$$V'(p) = \{f^k x : V(p, g^k y) = 1\}.$$

And let $M' = \langle X, f, V'\rangle$. Claim: $f^k x \in V'(B)$ iff $g^k y \models B$ for all $k \in \omega$ and formulas B in the language $L^{\bigcirc *}$. We prove this by induction on formulas.

Base case: For propositional variables p: $f^k x \in V'(p)$ iff $f^k x \in V'(p)$ iff $V(p, g^k y) = 1$ iff $g^k y \models p$.

Inductive step $\neg$, $\vee$: standard.

Inductive step $B = \bigcirc C$: $f^k x \in V'(\bigcirc C)$ iff $f^{k+1}x \in V'(C)$ iff $g^{k+1}y \models C$ (by IH) iff $g^k x \models \bigcirc C$.

Inductive step $B = *C$: $f^k x \in V'(*C)$ iff $(\forall n \geq k)(f^n x \in V'(C))$ iff $(\forall n \geq k)(g^n y \models C)$ (by IH) iff $g^k y \models *C$.

Thus $x \notin V'(A)$ since $y \not\models A$. So $A \notin \mathrm{DTL}^{\bigcirc *}_{\mathcal{F}}$, as desired. QED

COROLLARY 10.21 $\mathrm{DTL}^{\bigcirc *}_{0} = \mathrm{DTL}^{\bigcirc *}_{\mathcal{H}} = \mathrm{DTL}^{\bigcirc *}_{\mathcal{M}} = \mathrm{DTL}^{\bigcirc *}_{\mathcal{A}} = \mathrm{DTL}^{\bigcirc *}_{\mathbb{R}} = \mathrm{DTL}^{\bigcirc *}_{[0,1]} = \mathrm{DTL}^{\bigcirc *}_{\mathbb{R},\mathcal{H}} = \mathrm{DTL}^{\bigcirc *}_{\mathcal{A},\mathcal{H}} = \mathrm{DTL}^{\bigcirc *}_{fin} = \mathrm{LTL}$, *where fin is the class of finite topological spaces.*

5. S4 is topologically complete for (0, 1)

Here we combine ideas from several previous constructions into a short proof of topological completeness of the modal logic S4, first for the *binary* rational numbers (see below) in the interval (0, 1), and after that for real numbers in the same interval. Beginning with a finite, reflexive and transitive Kripke frame K, we present a direct definition of an open and continuous map from $(0, 1)$ onto the topological space corresponding to K. Given that S4 is complete w.r.t. finite, reflexive and transitive Kripke frames, this suffices for the completeness of S4 in (0, 1). The map we define is practically the same as in Mints and

Zhang, 2005a, very close to Aiello et al., 2003, and has some common features with Bezhanishvili and Gehrke, 2005.

Let us state a familiar general condition of propositional equivalence (more precisely, bisimulation) of two topological spaces. Recall that a map $f : X_1 \to X_2$ is *open* iff the image of any open set is open.

LEMMA 10.22 *Let X_1, X_2 be two topological spaces and f a continuous and open map from X_1 onto X_2. Let V_2 be a valuation for topological semantics on X_2 and define V_1 by the equation $V_1(p) = f^{-1}(V_2(p))$. Then for any formula A in the language $L^{\Box}$,*

$$\langle X_2, V_2\rangle \models A \text{ iff } \langle X_1, V_1\rangle \models A.$$

In particular, if A is refuted in X_2, then A is refuted in X_1.

Proof Induction on formulas. QED

Let

$$K = \langle W, R\rangle$$

be a finite Kripke frame with a reflexive and transitive accessibility relation R, a set of worlds $W = \{0, 1, \ldots, N\}$, and with root 0: $R0w$ holds for every $w \in W$. For every $w \in W$ the submodel with the root w is denoted by K_w. In particular $K_0 = K$. If w has no proper R-successors, that is Rww' implies $w' = w$, then w is a *leaf* of K. Speaking of *R-least*, *R-maximal* worlds we mean reading Rww' as $w \leq w'$. So the root 0 is the R-least element of W, a leaf is an R-maximal element. In general W may contain *clusters*, that is sets of R-equivalent elements. Every element of a cluster is R-least in the cluster.

Our plan is to define a continuous open function $\mathcal{W}$ from $\mathbf{Q}'$ onto K, where $\mathbf{Q}'$ is the set of binary rational numbers in the real interval $(0, 1)$:

$$\mathbf{Q}' = \{m/2^n : \ 0 < m < 2^n\}$$

5.1 Partitions

DEFINITION 10.23 *The* first partition $P_1(K)$, *corresponding to $K = \langle W, R\rangle$, of the open interval $(0, 1)$ of reals is defined as follows.*

If $N = 0$, that is the frame K has just one world 0, there are no endpoints and the whole interval (0,1) is marked K. All points $r \in \mathbf{Q}'$ are marked by the root 0: $\mathcal{W}_1(r) = 0$.

If $N > 0$, then the partition has endpoints

$$2^{-i}, \qquad i = 1, \ldots, 2N$$

that determine $2N + 1$ open intervals which are marked as follows:

$$\begin{aligned}(0, 2^{-2N}), (2^{-2N}, 2^{-2N+1}), \ldots,\\ (2^{-2i-1}, 2^{-2i}), (2^{-2i}, 2^{-2i+1}), \ldots, (2^{-1}, 1)\end{aligned} \tag{10.7}$$

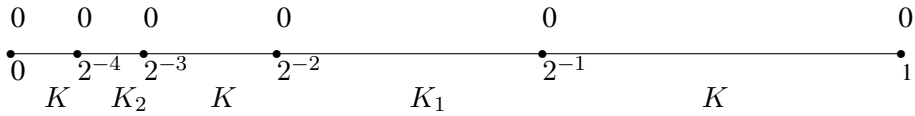


Figure 10.3. The first partition for a three-element model.

$$K, \qquad K_N, \qquad \ldots \qquad K, \qquad K_i, \qquad \ldots, \qquad K,$$

that is $\mathcal{W}_1((0, 2^{-2N})) = K, \ldots$. *The endpoints are marked by the root of the model* K*:*

$$\mathcal{W}_1(2^{-i}) = 0. \tag{10.8}$$

DEFINITION 10.24 *The* first partition $P_1((a,b),K)$ *of an arbitrary interval* (a,b), $a < b$ *corresponding to the frame* K *is proportional to (10.7). If* $N = 0$, *the whole interval is marked* K *and* $\mathcal{W}_1(r) = 0$ *for all* $r \in \mathbf{Q}'$.

If $N > 0$, *the partition has endpoints*

$$a + (b-a)2^{-i},\ i = 1, \ldots, 2N$$

that determine $2N+1$ *open intervals marked as in (10.7):*

$$K, \qquad K_{w_N}, \qquad \ldots K, \qquad K_{w_i}, \qquad \ldots, \qquad K \tag{10.9}$$

All endpoints are marked as in (10.8) by the root:

$$\mathcal{W}_1(a + (b-a)2^{-i}) = 0, \qquad i = 1, \ldots, 2N.$$

DEFINITION 10.25 ($(n+1)$-ST PARTITION) *Assume the* n*-th partition* $P_n(K)$ *of the interval* $(0,1)$ *corresponding to any finite frame* K *is already defined. If* $N = 0$, *define* $P_{n+1}(K) = P_n(K) = P_1(K)$.

Otherwise assume $P_n(K)$ *consists of endpoints*

$$a_1 < a_2 < \ldots < a_M$$

with each of the intervals (a_j, a_{j+1}) *marked by some* K_{w_j}, $w_j \in W$. *Then the* $(n+1)$*-st partition* $P_{n+1}(K)$ *of* $(0,1)$ *corresponding to* K *consists of all endpoints of* $P_n(K)$ *plus all endpoints of all partitions*

$$P_1((0,a_1),K), \ldots, P_1((a_i,a_{i+1}),K_{w_i}), \ldots, P_1((a_M,1),K). \tag{10.10}$$

Intervals of the new partition are marked according to partitions (10.10). The marking $\mathcal{W}_{n+1}(e_k) \in W$ *is an extension of* $\mathcal{W}_n$ *defined for new endpoints* e_i *according to partitions (10.10):*

$$\mathcal{W}_{n+1}(e_k) = \begin{cases} 0 & \text{if } e_k \in (0,a_1) \text{ or } e_k \in (a_M, 0) \\ w_j & \text{if } e_k \in (a_j, a_{j+1}). \end{cases}$$

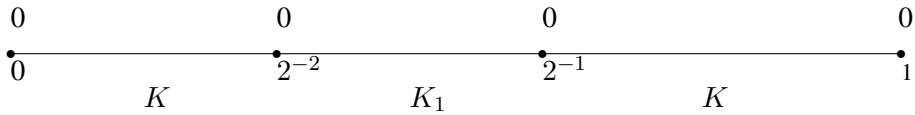


Figure 10.4. The first partition for a two-element model.

If K_{w_j} consists only of w_j, (that is w_j is a leaf of K), then $\mathcal{W}_{n+1}(r) = w_j$ for all $r \in (a_j, a_{j+1}) \cap \mathbf{Q}'$.

DEFINITION 10.26 *Define $P(K)$ as the union of all $P_n(K)$: the endpoints of $P(K)$ are all binary rational points $r \in \mathbf{Q}'$ of the interval $(0,1)$ marked as in the corresponding $P_n(K)$.*

$$\begin{aligned} \mathbf{n}(r) &= \text{the first number } n \text{ such that } r \text{ is an endpoint of } P_n(K) \text{ or } r \\ &\quad\ \text{belongs to an interval of } P_n(K) \text{ marked by } K_w \text{ for a leaf } w \\ \mathcal{W}(r) &= \mathcal{W}_{\mathbf{n}(r)}(r). \end{aligned}$$

EXAMPLE 10.27 *Consider a two-element Kripke frame $K = \langle \{0,1\}, \leq \rangle$:*

$$\begin{array}{cc} . & 1 \\ | & \\ . & 0 \end{array}$$

K is used to falsify the formula $(\Box p \vee \Box \neg p)$ by setting $V(p) = \{1\}$, so that

$$0 \not\models p, \ 1 \models p. \tag{10.11}$$

The first partition corresponding to the Kripke frame $K = \langle \{0,1\}, \leq \rangle$ is shown in Fig. 4. This shows that successive partitions are obtained by

1. *marking the endpoints by 0,*
2. *removing the second 1/4 of each of the remaining intervals and marking (all points in) the removed intervals by K_1; these intervals are not partitioned further,*
3. *marking the remaining intervals by $K = K_0$.*

Let P be the union of all intervals marked K_1 and $\bar{P} := (0,1) - P$. The valuation V leads to a topological model with the carrier $(0,1)$ and a valuation

$$V'(p) = P.$$

The set P is open, so $\mathbf{Int}(P) = P$, but $\mathbf{Int}(\bar{P}) = \emptyset$, since every binary rational point in $\bar{P}$ is a boundary of some interval of P. Hence

$$V'(\Box p \vee \Box \neg p) = \mathbf{Int}(P) \cup \mathbf{Int}(\bar{P}) = P \cup \emptyset = P \neq (0,1).$$

Hence $(\Box p \vee \Box \neg p)$ is not valid, as expected.

5.2 Properties of $P_n(K)$

The next lemma lists some combinatorial properties of partitions $P_n(K)$. The marking $w^j = w^j(r)$ in the clause 3 is determined by r and the partition $P_j(K)$.

LEMMA 10.28 *1 The functions $\mathbf{n}(r), \mathcal{W}(r)$ are defined for every $r \in \mathbf{Q}'$.*

2 For $m \geq \mathbf{n}(r)$, $\mathcal{W}_m(r) = \mathcal{W}_{\mathbf{n}(r)} = \mathcal{W}(r)$. If r is an endpoint of the partition $P_n(K)$, then one of the intervals of this partition adjacent to r is marked by $K_{\mathcal{W}(r)}$ and the other adjacent interval is marked by K_w with some w satisfying $R\mathcal{W}(r)w$.

For $\mathbf{n}(r) > 1$, r belongs to an interval of $P_{\mathbf{n}(r)-1}(K)$ marked by $K_{\mathcal{W}(r)}$.

3 For every $j < \mathbf{n}(r)$ the point r belongs to an interval I of the partition $P_j(K)$ marked by a K_{w^j} for a world w^j such that Rw^jw^{j+1} (where $w^{\mathbf{n}(r)-1} = \mathcal{W}(r)$).

4 If r belongs to an interval I of the partition $P_m(K)$ marked by a K_w, then $Rw\mathcal{W}(r)$.

Proof Induction on n. Note that if $r = k/2^n,\ 0 < k < 2^n$, then $\mathcal{W}_j(r)$ is defined beginning with $j = \mathbf{n}(r)$. QED

LEMMA 10.29 *For $r \in \mathbf{Q}'$*

(10.12) *$\mathcal{W}(r)$ is an R-least w such that $r = lim_{n\to\infty} r_n$*

for some sequence of $r_n \in \mathbf{Q}'$ with $r_n \neq r$ and $\mathcal{W}(r_n) = w$.

Proof Let $n = \mathbf{n}(r)$. If r belongs to an interval I of $P_n(K)$ marked by a leaf w of K, then $\mathcal{W}(r) = w = \mathcal{W}(r')$ for all $r' \in I$. Hence w is the only world satisfying (10.12).

If r is an endpoint of $P_n(K)$, then one of the the intervals of this partition adjacent to r (say from the left) is marked by $K_{\mathcal{W}(r)}$ and the other adjacent interval is marked by $K_{w'}$ with some w' satisfying $R\mathcal{W}(r)w'$. Hence $R\mathcal{W}(r)\mathcal{W}(r')$ holds for every r' in the union of these intervals, so that $r = lim_{n\to\infty} r_n$ and $\mathcal{W}(r_n) = w$ imply $R\mathcal{W}(r)w$. On the other hand, the left adjacent to r interval of any $P_m(K), m \geq n$ is marked $K_{\mathcal{W}}(r)$. If r_m is the midpoint of that interval, we have $\mathcal{W}_{m+1}(r_m) = \mathcal{W}(r) = 0$ as required. QED

Recall that the topology on K is determined by open sets K_w for $w \in W$.

LEMMA 10.30 *The mapping $\mathcal{W} : \mathbf{Q}' \to W$ is continuous: for every $r \in \mathbf{Q}'$ there is a $\delta > 0$ such that for every $r' \in \mathbf{Q}'$, if $|r - r'| < \delta$, then $\mathcal{W}(r') \in K_{\mathcal{W}(r)}$.*

Proof If $\mathcal{W}(r) = 0$, then $w \in K_{\mathcal{W}(r)} = K$ holds for all $w \in W$. Otherwise, take $n = \mathbf{n}(r) > 1$. Then r belongs to an interval I of $P_{n-1}(K)$ which was assigned $K_{\mathcal{W}(r)}$ and

$$\mathcal{W}(r') \in K_{\mathcal{W}(r)} \text{ for all } r' \in \mathbf{Q}' \cap I$$

as required. QED

LEMMA 10.31 *The mapping $\mathcal{W} : \mathbf{Q}' \to W$ is open: for every $r \in \mathbf{Q}'$, every $w' \in K_{\mathcal{W}(r)}$ and $\epsilon > 0$ there is an $r' \in \mathbf{Q}'$ with*

$$|r - r'| < \epsilon \text{ and } \mathcal{W}(r') = w'.$$

Proof Let $n = \mathbf{n}(r)$. Then $\mathcal{W}(r) = \mathcal{W}_{\mathbf{n}}(r)$ and for any $m \geq n$ the number r is an endpoint of an interval which is assigned $K_{\mathcal{W}(r)}$ in $P_m(K)$. Since $\mathcal{W}$ maps $\mathbf{Q}'$-points of this interval onto $K_{\mathcal{W}(r)}$, there is a point $r'_m \in \mathbf{Q}'$ in this interval such that $\mathcal{W}(r') = w'$. The sequence r'_m converges to r. QED

THEOREM 10.32 *$\mathbf{Q}'$ is complete for* S4.

Proof By Lemmas 10.30,10.31 and 10.22. QED

5.3 Extension to real numbers

The map $\mathcal{W}$ is extended here to real numbers in $(0, 1)$ by continuity using (10.12) as a hint, so that Lemmata 10.29, 10.30, 10.31 still hold. To make an R-least w in (10.12) unique, let's fix a *representative* $\rho(C) \in C$ for each cluster C of the model K.

DEFINITION 10.33 *For $x \in (0, 1) - \mathbf{Q}'$ define*

$$I(n, x) := \textit{the interval of the partition } P_n(K) \textit{ containing } x$$

$$\mathcal{W}(x) = \rho(C) \textit{ for the unique cluster } C \textit{ such that}$$

$$\exists n_0 (\forall n \geq n_0)(\exists w \in C)\ I(n, x) \textit{ is marked by } K_w \textit{ in } P_n(K). \tag{10.13}$$

LEMMA 10.34 *$\mathcal{W}(x)$ is defined for every $x \in (0, 1) - \mathbf{Q}'$.*

Proof Let $x \in (0, 1) - \mathbf{Q}'$ be fixed. Since $x \notin \mathbf{Q}'$, it is not an endpoint of $I(n, x)$ for any n. Let the world $M_n \in W$ be the marking of the interval $I(n, x)$ in $P_n(K)$, that is $I(n, x)$ is marked by K_{M_n}. For all n we have

$$RM_nM_{n+1}. \tag{10.14}$$

Indeed,

$$I(n+1,x) \subseteq I(n,x), \tag{10.15}$$

since for arbitrary intervals $I \in P_n(K),\ J \in P_{n+1}(K)$ either $I \supseteq J$ or I and J are disjoint. In the former case a relation

$$Rww'$$

for the marking K_w of I [in $P_n(K)$] and the marking $K_{w'}$ of J [in $P_{n+1}(K)$] follows from the definition of $P_{n+1}(K)$, hence (10.14).

Since the model K is finite, (10.14) implies that all worlds M_n are R-equivalent beginning with some n, that is $M_n \in C$ for one and the same cluster C, which is obviously unique. QED

LEMMA 10.35 *The function* $\mathcal{W} : (0,1) \to W$ *is continuous.*

Proof Similar to Lemma 10.30. Take an arbitrary $x \in (0,1)$. The case $W(x) = 0$ is obvious. Assume $\mathcal{W}(x) \neq 0$.

Case 1. $x \in \mathbf{Q}'$. Then $n(x) > 1$, since $\mathcal{W}(x) \neq 0$. For $n = n(x) > 1$ and $w = \mathcal{W}(x)$ the point x belongs to an interval of $P_{n-1}(K)$ which was marked by K_w. For all real numbers y in that interval $I(n-1,x)$ one has $Rw\mathcal{W}(y)$ by (10.14), as required.

Case 2. $x \notin \mathbf{Q}'$. Let $w = \mathcal{W}(x)$, that is for every $n \geq n_0$ the interval $I(n,x)$ is marked by $K_{w'}$ for some $w' \sim_R w$. Then $\mathcal{W}(y) \in K_w$ for all $y \in I(n_0,x)$, as required. QED

LEMMA 10.36 *The function* $\mathcal{W} : (0,1) \to W$ *is open.*

Proof For $x \in \mathbf{Q}'$ use Lemma 10.31.

Take $x \in (0,1) - \mathbf{Q}'$. For $n \geq n_0$ the point x belongs to an interval $I(n,x)$ marked by $K_{w'}$ for some $w' \sim_R \mathcal{W}(x)$. These intervals stabilize or converge to x and for each w' with Rww' each of them contains an $r \in \mathbf{Q}'$ with $\mathcal{W}(r) = w'$, as required. QED

THEOREM 10.37 *The interval* $(0,1)$ *is complete for* S4.

Proof By Lemmas 10.35, 10.36, 10.22. QED

6. The logic of homeomorphisms

Of particular interest is the class $\mathcal{H}$ of homeomorphisms (continuous bijections with continuous inverses). Intuitively, we keep track of *time* with f. Although our temporal modalities are forward-looking, it seems natural to keep track of time with functions that can look in both directions (i.e. that are bijective) and that are continuous in both directions. Despite the fact that our temporal modalities are forward-looking, restricting our attention to the class $\mathcal{H}$ makes a difference that can be expressed in our trimodal propositional language. In particular we have (10.16) and (10.17), below:

$$(\Box\bigcirc p \supset \bigcirc\Box p) \notin \mathrm{DTL}_0. \tag{10.16}$$

$$(\Box\bigcirc p \supset \bigcirc\Box p) \in \mathrm{DTL}_{\mathcal{H}}. \tag{10.17}$$

(10.17) follows from the fact that $Int(f^{-1}(S)) \subseteq f^{-1}(Int(S))$ where S is a subset of a topological space X on which f is a homeomorphism. To see (10.16), let $M = \langle X, f, V\rangle$ where $X = \{0,1\}$ with open sets $\emptyset$, $\{0\}$ and $\{0,1\}$; and where $f(0) = f(1) = 1$ and $V(p) = \{1\}$. The function f is continuous and hence M is a DTM. Also note that $V(\Box\bigcirc p) = \{0,1\}$ and $V(\bigcirc\Box p) = \emptyset$, so that $M \not\models (\Box\bigcirc p \supset \bigcirc\Box p)$.

As mentioned in Sec. 1, Konev et al. (2006a) present a proof that $\mathrm{DTL}_{\mathcal{H}}$ is not axiomatizable. We do, however, have an axiomatization of its next-interior fragment. Define the logic S4$\bigcirc$ as in Sec. 1. It turns out that S4$\bigcirc$ is complete for the class $\mathcal{H}$, i.e., $S4\bigcirc = \mathrm{DTL}_{\mathcal{H}}^{\bigcirc\Box}$. What's more, completeness holds for the real line $\mathbb{R}$ and similar spaces, providing an analogue to the McKinsey-Tarski-Kripke Theorem (Theorem 10.4).

THEOREM 10.38 (KREMER ET AL., 1997,KREMER AND MINTS, 1997) $\mathrm{S4}\bigcirc = \mathrm{DTL}_{\mathcal{H}}^{\bigcirc\Box} = \mathrm{DTL}_{\mathbb{R},\mathcal{H}}^{\bigcirc\Box} = \mathrm{DTL}_{[0,1],\mathcal{H}}^{\bigcirc\Box} = \mathrm{DTL}_{\mathcal{A},\mathcal{H}}^{\bigcirc\Box} = \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box} = \mathrm{DTL}_{\mathbb{R},\mathcal{O}}^{\bigcirc\Box} = \mathrm{DTL}_{[0,1],\mathcal{O}}^{\bigcirc\Box} = \mathrm{DTL}_{\mathcal{A},\mathcal{O}}^{\bigcirc\Box}$.

Proof (Vladimir Rybakov helped us with this proof.) The claim that $\mathrm{S4}\bigcirc \subseteq \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box}$ is just a version of soundness, which is proved as usual. Given this, the following inclusion relations are obvious:

$$\mathrm{S4}\bigcirc \subseteq \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathcal{H}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathbb{R},\mathcal{H}}^{\bigcirc\Box}$$
$$\mathrm{S4}\bigcirc \subseteq \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathbb{R},\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathbb{R},\mathcal{H}}^{\bigcirc\Box}$$
$$\mathrm{S4}\bigcirc \subseteq \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathcal{H}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{[0,1],\mathcal{H}}^{\bigcirc\Box}$$
$$\mathrm{S4}\bigcirc \subseteq \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{[0,1],\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{[0,1],\mathcal{H}}^{\bigcirc\Box}$$
$$\mathrm{S4}\bigcirc \subseteq \mathrm{DTL}_{\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathcal{A},\mathcal{O}}^{\bigcirc\Box} \subseteq \mathrm{DTL}_{\mathcal{A},\mathcal{H}}^{\bigcirc\Box}.$$

So it suffices to show, for every formula A in the language $L^{\bigcirc\Box}$:

if $[0,1], \mathcal{H} \models A$ then $\mathbb{R}, \mathcal{H} \models A$,
if $\mathbb{R}, \mathcal{H} \models A$ then $A \in$ S4$\bigcirc$, and
if $\mathcal{A}, \mathcal{H} \models A$ then $A \in$ S4$\bigcirc$.

See Theorems 10.39, 10.46 and 10.47, respectively. QED

THEOREM 10.39 *If* $[0,1], \mathcal{H} \models A$ *then* $\mathbb{R}, \mathcal{H} \models A$.

Proof Suppose that $\mathbb{R}, \mathcal{H} \not\models A$. Let $M = \langle \mathbb{R}, f, V \rangle$ be a model where f is a homeomorphism on $\mathbb{R}$ and where $M \not\models A$. Since f is a homeomorphism on $\mathbb{R}$, f is either strictly increasing or strictly decreasing. (In fact, as we show in the proof of Theorem 10.46, we can take f to be $f(x) = x + 1$. But we will continue with the more general case for now, since we have not yet shown Theorem 10.46.) Choose some strictly increasing continuous one-one function h from $\mathbb{R}$ onto the open interval $(0,1)$. Define f' on $[0,1]$ as follows:

$f'(x) = hfh^{-1}(x)$ if $0 < x < 1$
$f'(x) = x$ if f is strictly increasing and either $x = 0$ or $x = 1$
$f'(x) = 1 - x$ if f is strictly decreasing and either $x = 0$ or $x = 1$.

And define

$$V'(p) = \{x \in (0,1) : h^{-1}(x) \in V(p)\}.$$

f' is one-one and onto. f' is also continuous. For if f is strictly increasing then $lim_{x \to 0} f'(x) = 0$ and $lim_{x \to 1} f'(x) = 1$; and if f is strictly decreasing then $lim_{x \to 0} f'(x) = 1$ and $lim_{x \to 1} f'(x) = 0$. So $M' = \langle [0,1], f', V' \rangle$ is a dynamic topological model.

Notice that $(0,1) \cap V'(B) = \{x \in (0,1) : h^{-1}(x) \in V(B)\}$, for every formula B. The proof of this is a routine induction on formulas. So $V'(A) \neq [0,1]$. For otherwise we would have $V(A) = \mathbb{R}$, which is false. So $[0,1], \mathcal{H} \not\models A$, as desired. QED

Before we prove Theorems 10.46 and 10.47, some definitions and lemmas.

DEFINITION 10.40 *Given a formula B, let $g(B)$ be the result of pushing all the occurrences of $\bigcirc$ to the atomic formulas. For example, $g(\bigcirc(\bigcirc\Box(p \vee \bigcirc q) \vee \bigcirc\neg r)) = (\Box(\bigcirc\bigcirc p \vee \bigcirc\bigcirc\bigcirc q) \vee \neg\bigcirc\bigcirc r)$. To be more precise, define $g(B)$ inductively as follows:*

$g(\bigcirc^n B) = \bigcirc^n B$, *if* $B \in PV$,

$g(\bigcirc^n \neg B) = \neg g(\bigcirc^n B)$,
$g(\bigcirc^n (B \vee C)) = g(\bigcirc^n B) \vee g(\bigcirc^n C)$, *and*
$g(\bigcirc^n \Box B) = \Box g(\bigcirc^n B)$.

DEFINITION 10.41 *A* near-atom *is a formula of the form* $\bigcirc^n p$ *where* $p \in PV$.

DEFINITION 10.42 *A formula is* simple *iff it is built up from near-atoms using the Boolean connectives and* $\Box$. *Simple formulas are the formulas in the range of* g.

CONVENTION 10.43 We will take S4 to be formulated by its standard axioms and rules, for a language whose formulas are just the simple formulas, treating the near atoms as indivisible atomic formulas. We also slightly restate the definition of *topological model*, Definition 10.1: A topological model now becomes an ordered pair, $M = \langle X, V \rangle$, where X is a topological space and V assigns a subset of X *to each near atom* $\bigcirc^n p$ rather than to each propositional variable p. Mimicking Definition 10.1, we extend V to all *simple* formulas as follows:

$V(\bigcirc^n p) = V(\bigcirc^n p)$,
$V(A \vee B) = V(A) \cup V(B)$,
$V(\neg B) = X - V(B)$, and
$V(\Box B) = Int(V(B))$.

As in Definition 10.1, We define standard validity relations:

$M \models B$ iff $V(B) = X$.
$X \models B$ iff $M \models B$ for every model $M = \langle X, V \rangle$.
B is *valid* ($\models B$) iff $X \models B$ for every topological space X.

The McKinsey-Tarski-Kripke Theorem (Theorem 10.4) still holds: Suppose that X is a dense-in-itself metric space and A is a simple formula. Then the following are equivalent: (i) $A \in S4$; (ii) $\models A$; (iii) $X \models A$; (iv) $\mathbb{R} \models A$; (v) $Y \models A$ for every finite topological space Y; and (vi) $Y \models A$ for every Alexandrov space Y.

LEMMA 10.44 $B \in$ S4$\bigcirc$ *iff* $g(B) \in$ S4 *iff* $g(B) \in$ S4$\bigcirc$.

Proof By a standard induction on the proof of B in S4$\bigcirc$, we can show that if $B \in$ S4$\bigcirc$ then $g(B) \in$ S4. It is obvious that if $g(B) \in$ S4 then $g(B) \in$ S4$\bigcirc$. Finally, if $g(B) \in$ S4$\bigcirc$ then $B \in$ S4$\bigcirc$, since $(B \equiv g(B)) \in$ S4$\bigcirc$. QED

LEMMA 10.45 *For every formula* B, $g(B) \in$ S4 *iff* $(0,1) \models g(B)$ *where* $(0,1)$ *is the open unit interval.*

Proof This follows from Theorem 10.4 and Lemma 10.44. QED

THEOREM 10.46 *If* $\mathbb{R}, \mathcal{H} \models A$ *then* $A \in$ S4$\bigcirc$.

Proof Suppose that $A \notin$ S4$\bigcirc$. Then, by Lemmas 10.44 and 10.45, for some topological model $M = \langle (0,1), V \rangle$, we have $M \not\models g(A)$. Let M' be the dynamic topological model $\langle \mathbb{R}, f, V' \rangle$, where $fx = x+1$ and $V'(p) = \{x \in \mathbb{R}$: for some natural number m, $x - m \in V(\bigcirc^m p)\}$. See Fig. 10.5. f is a

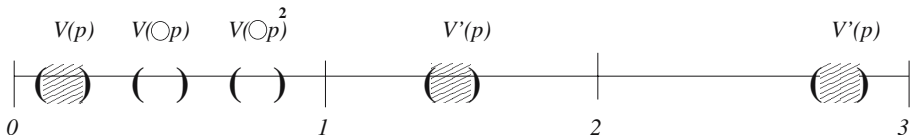


Figure 10.5. Definition of V'.

homeomorphism. We will be done if we can show that $M' \not\models A$. For this, it suffices to show that $M' \not\models g(A)$, because of Lemma 10.44 and because of soundness. And for this it suffices to show that for every simple formula B, we have $V(B) = (0,1) \cap V'(B)$. We show this by induction on the construction of B.

Base case: B is a near atom, say $\bigcirc^n p$. Note the following:
$x \in (0,1) \cap V'(B)$
$\Rightarrow x \in (0,1)$ and $x \in V'(\bigcirc^n p)$
$\Rightarrow x \in (0,1)$ and $x + n \in V'(p)$
$\Rightarrow x \in (0,1)$ and, for some m, $x + n - m \in V(\bigcirc^m p)$
$\Rightarrow m = n$, since $x \in (0,1)$ and $x + n - m \in V(\bigcirc^m p) \subseteq (0,1)$
$\Rightarrow x \in (0,1)$ and $x \in V(\bigcirc^n p)$
$\Rightarrow x \in V(B)$.

Conversely, $x \in V(B)$
$\Rightarrow x \in V(B)$
$\Rightarrow x \in (0,1)$ and $x \in V(\bigcirc^n p)$
$\Rightarrow x \in (0,1)$ and, for some m, $x + n - m \in V(\bigcirc^m p)$
$\Rightarrow x \in (0,1)$ and $x + n \in V'(p)$
$\Rightarrow x \in (0,1)$ and $x \in V'(\bigcirc^n p)$
$\Rightarrow x \in (0,1) \cap V'(B)$.

Inductive step $B = C \vee D$. $V(C \vee D) = V(C) \cup V(D) = ((0,1) \cap V'(C)) \cup ((0,1) \cap V'(D)) = (0,1) \cap V'(C \vee D)$.

Inductive step $B = \neg C$. $V(\neg C) = (0,1) - V(C) = (0,1) - ((0,1) \cap V'(C)) = (0,1) - (\mathbb{R} \cap V'(C)) = (0,1) \cap V'(\neg C)$.

Inductive step $B = \Box C$. $V(\Box C) = Int(V(C)) = Int((0,1) \cap V'(C)) = Int((0,1)) \cap Int(V'(C)) = (0,1) \cap V'(\Box C)$. QED

THEOREM 10.47 *If* $\mathcal{A}, \mathcal{H} \models A$ *then* $A \in$ S4$\bigcirc$.

Proof Suppose that $A \notin$ S4$\bigcirc$. Then $g(A) \notin S4$. So there is a Kripke model $M = \langle W, R, V \rangle$ (where $\langle W, R \rangle$ is a Kripke frame) such that $M \not\models g(A)$. Now define a dynamic topological model $M' = \langle X, f, V' \rangle$ as follows:

$X = \{\langle w, n \rangle : w \in W$ and n is an integer$\}$,
$\langle w, n \rangle R' \langle w', m \rangle$ iff wRw' and $n = m$,
$Y \subseteq X$ is *open* iff Y is closed under the relation R',
$f\langle w, n \rangle = \langle w, n+1 \rangle$, and
$\langle w, n \rangle \in V'(p)$ iff $w \in V(\bigcirc^n p)$.

X is a topological space, if we take the topology of *open* sets as defined directly above. In fact, X is an Alexandrov space (see Definition 10.3). f is both continuous and open since $\langle w, n \rangle R' \langle w', m \rangle$ iff $f\langle w, n \rangle R' f\langle w', m \rangle$. And f is clearly one-one and onto. So $M' = \langle X, f, V' \rangle$ is a dynamic Alexandrov model, with f a homeomorphism. We will be done if we can show that $M' \not\models A$. For this, it suffices to show that $M' \not\models g(A)$, because of Lemma 10.44 and because of soundness. And for this it suffices to show that for every simple formula B and every $w \in W$ we have $w \in V(B)$ iff $\langle w, 0 \rangle \in V'(B)$. We show this by induction on the construction of B.

Base case: B is a near atom, say $\bigcirc^n p$. Then $\langle w, 0 \rangle \in V'(B)$ iff $\langle w, 0 \rangle \in V'(\bigcirc^n p)$ iff $f^n\langle w, 0 \rangle \in V'(p)$ iff $\langle w, n \rangle \in V'(p)$ iff $\langle w, n \rangle \in V'(p)$ iff $w \in V(\bigcirc^n p)$ iff $w \in V(\bigcirc^n p)$ iff $w \in V(B)$.

Inductive step $B = C \vee D$. $\langle w, 0 \rangle \in V'(C \vee D)$ iff $\langle w, 0 \rangle \in V'(C)$ or $\langle w, 0 \rangle \in V'(D)$ iff $w \in V(C)$ or $w \in V(D)$ iff $w \in V(C \vee D)$.

Inductive step $B = \neg C$. $\langle w, 0 \rangle \in V'(\neg C)$ iff $\langle w, 0 \rangle \notin V'(C)$ iff $w \notin V(C)$ iff $w \in V(\neg C)$.

Inductive step $B = \Box C$. $\langle w, 0 \rangle \in V'(\Box C)$ iff $(\forall w')(\forall n)$(if $\langle w, 0 \rangle R' \langle w', n \rangle$ then $\langle w', n \rangle \in V'(C)$) iff $(\forall w')$(if (wRw' then $\langle w', 0 \rangle \in V'(C)$) iff $(\forall w')$(if wRw' then $w' \in V(C)$) iff $w \in V(\Box C)$. QED

THEOREM 10.48 *If* $fin, \mathcal{H} \models A$ *then* $A \in$ S4$\bigcirc$, *where* fin *is the class of* finite *topological spaces.*

Proof Suppose that $A \notin$ S4$\bigcirc$. Then, by Lemma 10.44 and the finite model property for S4 we have $M \not\models g(A)$, for some topological model $M = \langle X, V \rangle$

where X is finite. Let n be the maximum number of consecutive occurrences of $\bigcirc$ in the formula $g(A)$. For each $k = 0, \ldots, n$, define the finite topological space X_k as follows:

$$X_k = \{\langle x, k\rangle : x \in X\},$$

where a set $O \subseteq X_k$ is *open* in X_k iff the set $\{x : \langle x, k\rangle \in O\}$ is open in X. Define the function $f_k : X \to X_k$ as follows: $f_k(x) = \langle x, k\rangle$. f_k is clearly a homeomorphism from X onto X_k.

For $S \subseteq X$, define $Int_X(S)$ as the interior of S in X. Similarly, for $S \subseteq X_k$ define $Int_{X_k}(S)$ as the interior of S in X_k, and for $S \subseteq X'$ define $Int_{X'}(S)$ as the interior of S in X'. Then note that, for any $S \subseteq X'$, we have $X_k \cap Int_{X'}(S) = Int_{X_k}(X_k \cap S)$. Also note that, for any $S \subseteq X$, we have $Int_X(S) = f_0^{-1}(Int_{X_0}(f_0(S)))$.

Next, define the topological space $X' = \cup_k X_k$, where a set O is open in X' iff the following set is open in X_k for each k: $O \cap X_k$. Define the function $f : X \to X$ as follows:

$$f(\langle x, k\rangle) \quad = \quad \begin{cases} \langle x, k+1\rangle, \text{ if } k < n \\ \langle x, 0\rangle, \text{ if } k = n. \end{cases}$$

f is clearly a homeomorphism from X onto X. Finally, define the valuation function $V' : PV \to X'$ as follows: $V'(p) = \{\langle x, k\rangle : x \in V(\bigcirc^k p)\}$.

Let M' be the dynamic topological model $\langle X', f, V'\rangle$. We will be done if we can show that $M' \not\models A$. For this, it suffices to show that $M' \not\models g(A)$, because of Lemma 10.44 and because of soundness. And for this it suffices to show that for every simple formula B with n or fewer consecutive occurrences of $\bigcirc$, we have

$$V(B) = \{x \in X : \langle x, 0\rangle \in V'(B)\}.$$

We show this by induction on the construction of B.

Base case: B is a near atom, say $\bigcirc^k p$, where $k \leq n$. Note: $\langle x, 0\rangle \in V'(B) \Leftrightarrow \langle x, 0\rangle \in V'(\bigcirc^k p) \Leftrightarrow f^k(\langle x, 0\rangle) \in V'(p) \Leftrightarrow \langle x, k\rangle \in V'(p) \Leftrightarrow x \in V(\bigcirc^k p) \Leftrightarrow x \in V(B)$.

Inductive step $B = C \vee D$. Note: $\langle x, 0\rangle \in V'(C \vee D) \Leftrightarrow \langle x, 0\rangle \in V'(C)$ or $\langle x, 0\rangle \in V'(D) \Leftrightarrow x \in V(C)$ or $x \in V(D) \Leftrightarrow x \in V(C \vee D)$.

Inductive step $B = \neg C$. Note: $\langle x, 0\rangle \in V'(\neg C) \Leftrightarrow \langle x, 0\rangle \notin V'(C) \Leftrightarrow x \notin V(C) \Leftrightarrow x \in V(\neg C)$.

Inductive step $B = \Box C$. Note: $\langle x, 0\rangle \in V'(\Box C)$
$\Leftrightarrow \langle x, 0\rangle \in Int_{X'}(V'(C))$
$\Leftrightarrow \langle x, 0\rangle \in X_0 \cap Int_{X'}(V'(C))$
$\Leftrightarrow \langle x, 0\rangle \in Int_{X_0}(X_0 \cap V'(C))$
$\Leftrightarrow \langle x, 0\rangle \in Int_{X_0}(\{\langle y, 0\rangle : \langle y, 0\rangle \in V'(C)\})$
$\Leftrightarrow \langle x, 0\rangle \in Int_{X_0}(\{\langle y, 0\rangle : y \in V(C)\})$, by the inductive hypothesis
$\Leftrightarrow f_0(x) \in Int_{X_0}(f_0(V(C)))$

$\Leftrightarrow x \in f_0^{-1}(Int_{X_0}(f_0(V(C))))$
$\Leftrightarrow x \in Int_X(V(C))$
$\Leftrightarrow x \in V(\Box C)$. QED

COROLLARY 10.49 S4$\bigcirc$ = DTL$_{fin,\mathcal{H}}$.

7. The logic of continuous functions

The most basic dynamic topological logic is the logic of continuous functions on topological spaces, i.e. DTL$_0 = \{A : \models A\}$. It is not known whether DTL$_0$ is axiomatizable. In this section, we will prove that the next-interior fragment of DTL$_0$ *is* axiomatizable: it is axiomatized by the system S4C, defined in Sec. 1, above. (See Sec. 7.1.) Thus, S4C is the most general next-interior logic of continuous functions on topological spaces. This is a corollary to our theorem that S4C is the next-interior logic of all dynamic Kripke frames (Sec. 7.1). We also prove that S4C satisfies the finite model property (Sec. 7.2), and that S4C is the next-interior logic of continuous functions on Cantor space (Sec. 7.5).

Together, these results provide a partial analogue to the McKinsey-Tarski-Kripke theorem (Theorem 10.4), above: For every formula A of $L^{\bigcirc\Box}$, we have $A \in$ S4C iff $\models A$ iff $Y \models A$ for every Kripke model Y iff $Y \models A$ for every finite Kripke model Y. To this extent, the situation with S4C is similar to the situation with S4$\bigcirc$ (see Theorem 10.38, above.) But there is an important disanalogy: though S4$\bigcirc$ is the next-interior logic of homeomorphisms on $\mathbb{R}$, S4C is *not* the next-interior logic of continuous functions on $\mathbb{R}$. Presently, we will give an example of a formula A in the language $L^{\bigcirc\Box}$ such that $A \notin$ S4C but $\mathbb{R} \models A$. See Slavnov, 2005 for another example. We note that the axiomatizability of the next-interior logic of the real line remains an open problem.

Consider the following formula A, where p and q are propositional variables:

$$(\Box\bigcirc p \supset \bigcirc\Diamond\Box p) \vee (\bigcirc q \supset \Box\bigcirc q).$$

We first show that $A \notin$ S4C. Let $M = \langle X, f, V\rangle$, where

X = { 0, 1, 2 };
the open sets are $\emptyset$, X, and $\{2\}$;
$f(2) = f(1) = 0$ and $f(0) = 1$; and
$V(p) = \{0, 1\}$, and $V(q) = \{1\}$.

Note the following:

$V(\bigcirc p) = X$; so $V(\Box\bigcirc p) = X$.
$V(\Box p) = \emptyset$; so $V(\bigcirc\Diamond\Box p) = \emptyset$.
Thus $V(\Box\bigcirc p \supset \bigcirc\Diamond\Box p) = \emptyset$.

Meanwhile, $V(\bigcirc q) = \{0\}$; so $V(\Box\bigcirc q) = \emptyset$.
Thus $V(\bigcirc q \supset \Box\bigcirc q) = \{1, 2\}$.
Thus $V(A) = \{1, 2\} \neq X$.
Thus $M \not\models A$.

We now show that $\mathbb{R} \models A$. Suppose not. Then there is some dynamic topological model $M' = \langle \mathbb{R}, f', V' \rangle$ and some $x \in \mathbb{R}$ such that $x \notin V'(A)$. Thus,

(i) $x \in V'(\Box\bigcirc p)$. So there is an open interval I such that $x \in I \subseteq V'(\bigcirc p)$. So $f'(x) \in f'(I) \subseteq V'(p)$.

(ii) $x \notin V'(\bigcirc\Diamond\Box p)$. So $f'(x) \notin Cl(Int(V'(p)))$.

(iii) $x \in V'(\bigcirc q)$. So $f'(x) \in V(q)$.

(iv) $x \notin V'(\Box\bigcirc q)$. So there is some $y \in I$ such that $y \notin V'(\bigcirc q)$. Thus $f'(y) \notin V'(q)$. Thus $f'(x) \neq f'(y)$. Thus $f'(I)$ is not a singleton set.

Since $f'(I)$ is not a singleton set and since I is an open interval, $f'(I)$ is either an open interval, a closed interval, or a semi-closed interval, i.e. an interval of the form $[a, b)$ or $(a, b]$. In any case, $f'(I) \subseteq Cl(Int(f'(I)))$. And since from (i) we have $f'(I) \subseteq V'(p)$, we also have

$$f'(x) \in f'(I) \subseteq Cl(Int(f'(I))) \subseteq Cl(Int(V'(p))).$$

But this contradicts (ii).

7.1 Canonical models

We begin with the completeness of S4C for Kripke models. Recall some standard notions: A is a *theorem* iff $A \in$ S4C. A is *consistent* iff $\neg A \notin$ S4C. A *theory* is a set of formulas in the language $L^{\bigcirc\Box}$ containing all the theorems of S4C and closed under Modus Ponens. A theory T is *complete* iff for every formula A either $A \in T$ or $\neg A \in T$. A theory T is *consistent* iff some formula is not in T. A set S of formulas is *consistent* iff some theory $T \supseteq S$ is consistent.

THEOREM 10.50 (ARTEMOV ET AL., 1997 AND DAVOREN, 1998) S4C *is sound and complete for the class of all dynamic Kripke models (and hence for all dynamic Alexandrov models).*

Proof Soundness is obvious. For completeness it suffices to construct a canonical dynamic Kripke model M (see Definition 10.10) such that $M \models A$ iff $A \in$ S4C, for every formula A in the language $L^{\bigcirc\Box}$. In fact, given soundness, it will suffice to show that if $M \models A$ then $A \in S4C$.

Define a Kripke frame $\langle X, R \rangle$ and a function f on X as follows:

$$X = \{x : x \text{ is a complete consistent theory}\};$$

xRy iff for every formula A, if $\Box A \in x$ then $A \in y$; and

$$fx = \{A : \bigcirc A \in x\}.$$

Note that R is reflexive since $(\Box A \supset A) \in$ S4C and transitive since $(\Box A \supset \Box\Box A) \in$ S4C.

Now we show that f is *monotone* : $xRy \Rightarrow (fx)R(fy)$. So suppose that xRy. To see that $(fx)R(fy)$, suppose $\Box A \in fx$. Then $\bigcirc\Box A \in x$. So $\Box\bigcirc A \in x$, since $(\bigcirc\Box A \supset \Box\bigcirc A) \in$ S4C. So $\bigcirc A \in y$. So $A \in fy$, as desired.

Thus $\langle X, f\rangle$ is a dynamic Kripke system. Define $V(p) = \{x \in X : p \in x\}$. Then $M = \langle X, f, V\rangle$ is a dynamic Kripke model. By a standard induction on the complexity of the formula A, we have $x \in V(A)$ iff $A \in x$, for every $x \in X$.

To show that if $M \models A$ then $A \in$ S4C, suppose that $A \notin$ S4C. Then $\neg A$ is consistent. By a standard argument, every consistent formula is a member of some complete consistent theory. So $\neg A \notin x$, for some $x \in X$. So $x \notin V(A)$. So $M \not\models A$, as desired. QED

COROLLARY 10.51 S4C *is sound and complete for the class of all dynamic topological systems.*

7.2 The finite model property for S4C

We can improve on the last corollary as follows:

THEOREM 10.52 S4C *is sound and complete for the class of all finite rooted dynamic topological systems.*

In particular, we will show that if $A \notin$ S4C, then there is some finite rooted dynamic Kripke model $M = \langle W,\ R,\ f,\ V\rangle$ such that $M \not\models A$.

Let a *signed formula* be any ordered pair $\langle \pm, A\rangle$ where A is a formula of $L^{\bigcirc\Box}$. We will write $+A$ for $\langle +, A\rangle$ and $-A$ for $\langle -, A\rangle$. A *pseudo-atom* is a set of signed formulas. Given a pseudo-atom α, let $|\alpha| = \{A : \pm A \in \alpha\}$. We identify any nonempty pseudo-atom α with its conjunction as follows: we identify $\{+A, -B, -C\}$ with $A \mathbin{\&} \neg B \mathbin{\&} \neg C$. We identify the pseudo-atom $\emptyset$ with the formula $p \vee \neg p$. We say that a formula A is *consistent* just in case $\neg A \notin$ S4C. A pseudo-atom is *consistent* just in case the formula with which it is identified is consistent.

We say that a set S of formulas is *strongly closed* iff S satisfies the following closure conditions, for any formulas B and C, and for $n, m \geq 0$:

(CC1) if $\bigcirc^{n+1}B \in S$ then $\bigcirc^n B \in S$
(CC2) if $\bigcirc^n\Box B \in S$ then $\bigcirc^n B \in S$,

(CC3) if $\bigcirc^n \neg B \in S$ then $\bigcirc^n B \in S$,
(CC4) if $\bigcirc^n (B \,\&\, C) \in S$ then $\bigcirc^n B \in S$ and $\bigcirc^n C \in S$,
(CC5) if $\bigcirc^n (B \vee C) \in S$ then $\bigcirc^n B \in S$ and $\bigcirc^n C \in S$, and
(CC6) if $\bigcirc^n \bigcirc^{m+1} B \in S$ then $\bigcirc^n \Box \bigcirc^{m+1} B \in S$.

Note that if S satisfies (CC1)-(CC5) then S is closed under subformulas. We say that a pseudo-atom α is *strongly closed* just in case the set $|\alpha|$ of formulas is strongly closed. If S is a strongly closed set of formulas, then an *S-atom* is any consistent strongly closed pseudo-atom α with $|\alpha| \subseteq S$.

Suppose that the formula $A \notin$ S4C. We will define a *finite rooted* dynamic Kripke model M such that $M \not\models A$. Let S be the smallest strongly closed set of formulas containing A. Note that S is finite. Define the dynamic Kripke model $M = \langle W,\ R,\ f,\ V \rangle$ as follows:

1 W = the set of S-atoms.

2 $\alpha R \beta$ iff, for every formula B, if $+\Box B \in \alpha$ then $+\Box B \in \beta$.

3 $f(\alpha) = \{+B : + \bigcirc B \in \alpha\} \cup \{-B : -\bigcirc B \in \alpha\}$.

4 $V(p) = \{\alpha : +p \in \alpha\}$, for each propositional variable p.

We have to make sure that M is a dynamic Kripke model. It is obvious that R is reflexive and transitive. We must still show two things:

(1) If $\alpha \in W$ then $f\alpha \in W$, i.e. if α is an S-atom then $f\alpha$ is an S-atom; and
(2) f is monotonic.

We will show these below. Given (1) and (2), M is a dynamic Kripke model. Also note that M is finite, since S is finite. Also, M is rooted: $\emptyset$ is an S-atom; for every S-atom α we have $\emptyset R \alpha$; and $f(\emptyset) = \emptyset$. Below, we will also show the following, for every formula B:

(3) For every S-atom α, if $\pm B \in \alpha$ then ($\alpha \in V(B)$ iff $+B \in \alpha$).

Given (3), we are just about done. Since $A \notin$ S4C, there is some S-atom α with $-A \in \alpha$. Thus $+A \notin \alpha$. Thus $\alpha \notin V(A)$, by (3). Thus $M \not\models A$.

All that remains for Theorem 10.52 is to prove (1), (2) and (3).

Proof of (1). Suppose that α is an S-atom. We must show that (1.1) $f\alpha$ is consistent, (1.2) $|f\alpha| \subseteq S$ and (1.3) $f\alpha$ is strongly closed. If $f\alpha$ is empty, then (1.1) is satisfied since $f\alpha$ is identified with the consistent formula $p \vee \neg p$; and (1.2) and (1.3) are trivially satisfied. So we assume that $f\alpha$ is nonempty. Re (1.1): $f\alpha$ is inconsistent $\Rightarrow \bigcirc f\alpha$ is inconsistent $\Rightarrow \alpha$ is inconsistent since $(\alpha \supset \bigcirc f\alpha) \in$ S4C. Re (1.2): Note that $|f\alpha| \subseteq |\alpha|$, since $|\alpha|$ is closed under subformulas; thus $|f\alpha| \subseteq S$ since $|\alpha| \subseteq S$. Re (1.3), we must show that $|f\alpha|$

satisfies the closure conditions (CC1)-(CC6) above. Re (CC1): Suppose that $\bigcirc^{n+1}B \in |f\alpha|$. Then $\bigcirc^{n+2}B \in |\alpha|$. So $\bigcirc^{n+1}B \in |\alpha|$, since $|\alpha|$ satisfies (CC2). So $\bigcirc^{n}B \in |f\alpha|$, by the definition of f. Re (CC2): Suppose that $\bigcirc^{n}\Box B \in |f\alpha|$. Then $\bigcirc^{n+1}\Box B \in |\alpha|$. So $\bigcirc^{n+1}B \in |\alpha|$, since $|\alpha|$ satisfies (CC2). So $\bigcirc^{n}B \in |f\alpha|$, by the definition of f. Similarly for (CC3)-(CC5). Re (CC6): Suppose that $\bigcirc^{n}\bigcirc^{m+1}B \in |f\alpha|$. Then $\bigcirc^{n+1}\bigcirc^{m+1}B \in |\alpha|$. So $\bigcirc^{n+1}\Box\bigcirc^{m+1}B \in |\alpha|$, since $|\alpha|$ satisfies (CC6). So $\bigcirc^{n}\Box\bigcirc^{m+1}B \in |f\alpha|$.

Proof of (2). Suppose that $\alpha R\beta$: we want to show that $f\alpha\, R\, f\beta$. So suppose that $+\Box A \in f\alpha$. Then $+\bigcirc\Box A \in \alpha$. Since α is strongly closed, $|\alpha|$ satisfies (CC6), above; thus either $+\Box\bigcirc\Box A \in \alpha$ or $-\Box\bigcirc\Box A \in \alpha$. Note that $\bigcirc\Box A \supset \Box\bigcirc\Box A \in$ S4C. So $+\Box\bigcirc\Box A \in \alpha$, by the consistency of α. Thus $+\Box\bigcirc\Box A \in \beta$, since $\alpha R\beta$. Thus $+\bigcirc\Box A \in \beta$, by the strong closure and the consistency of β. Thus $+\Box A \in f\beta$, as desired.

Proof of (3). We proceed by induction.

Base case, B is a propositional variable. Cf the definition of V.

Inductive step, $B = C \,\&\, D$. Suppose that $\pm B \in \alpha$. Note that $\alpha \in V(B)$ iff $\alpha \in V(C \,\&\, D)$ iff $\alpha \in V(C)$ and $\alpha \in V(D)$ iff $+C \in \alpha$ and $+D \in \alpha$ iff $+B \in \alpha$, by the consistency of α.

Inductive step, $B = C \vee D$. Suppose that $\pm B \in \alpha$. Note that $\alpha \in V(B)$ iff $\alpha \in V(C \vee D)$ iff $\alpha \in V(C)$ or $\alpha \in V(D)$ iff $+C \in \alpha$ or $+D \in \alpha$ iff $+B \in \alpha$, by the consistency of α.

Inductive step, $B = \neg C$. Suppose that $\pm B \in \alpha$. Note that $\alpha \in V(B)$ iff $\alpha \notin V(C)$ iff $+C \notin \alpha$ iff $-C \in \alpha$ iff $+B \in \alpha$, by the consistency of α.

Inductive step, $B = \bigcirc C$. Suppose that $\pm B \in \alpha$. Note that $\alpha \in V(B)$ iff $f\alpha \in V(C)$ iff $+C \in f\alpha$ iff $+\bigcirc C \in \alpha$ iff $+B \in \alpha$.

Inductive step, $B = \Box C$. Suppose that $\pm B \in \alpha$. We consider the directions of the biconditional separately.

($\Rightarrow$) Suppose that $+B \notin \alpha$. Let $\gamma = \{+\Box D : +\Box D \in \alpha\}$ and let $\delta = \gamma \cup \{-C\}$. First note that δ is a consistent pseudo-atom: δ is inconsistent $\Rightarrow$ $(\gamma \supset C) \in$ S4C $\Rightarrow$ $(\gamma \supset \Box C) \in$ S4C (since γ is a conjunction of formulas of the form $\Box D$) $\Rightarrow$ α is inconsistent. Since δ is consistent and $|\delta| \subseteq S$, there is an S-atom β such that $\delta \subseteq \beta$. Since $-C \in \beta$, we have $C \notin \beta$. Thus, by the inductive hypothesis, $\beta \notin V(C)$. Also note that $\alpha R\beta$. Thus $\alpha \notin V(\Box C) = V(B)$.

($\Leftarrow$): Suppose that $+B \in \alpha$. To show that $\alpha \in V(B)$, consider any S-atom β with $\alpha R\beta$. Note that $+\Box C = +B \in \beta$ by the definition of R. So $+C \in \beta$, by the consistency of β. So, by the inductive hypothesis, $\beta \in V(C)$. Thus $\alpha \in V(\Box C) = V(B)$, as desired.

7.3 Bisimulation of Dynamic Topological Systems

DEFINITION 10.53 *Let $M_1 = \langle X_1, T_1 \rangle$, $M_2 = \langle X_2, T_2 \rangle$ be two dynamic topological spaces. We say a map $\mathcal{W} : M_1 \to M_2$ is* commuting *if*

1 *$\mathcal{W}$ is a continuous and open map from X_1 onto X_2, and*

2 *$\mathcal{W}(T_1(x)) = T_2(\mathcal{W}(x))$.*

LEMMA 10.54 *Let $M_1 = \langle X_1, T_1, V_1 \rangle, M_2 = \langle X_2, T_2, V_2 \rangle$ be two dynamic topological models. Suppose that $\mathcal{W} : M_1 \to M_2$ is a commuting map and for each propositional variable p,*

$$V_1(p) = \mathcal{W}^{-1}(V_2(p)).$$

Then

$$V_1(A) = \mathcal{W}^{-1}(V_2(A))$$

for any formula A of $L^{\bigcirc\Box}$.

Proof By induction on A. The base case and induction steps for connectives $\vee, \wedge, \neg$ are straightforward. Now consider the remaining two cases: $A \equiv \Box B$ and $A \equiv \bigcirc B$.

- Case $A \equiv \Box B$. We have

$$\begin{array}{rcll} V_1(A) & = & V_1(\Box B) & \\ & = & Int(V_1(B)) & \textit{by the definition of } V_1 \\ & = & Int(\mathcal{W}^{-1}(V_2(B))) & \textit{by the induction hypothesis} \\ & = & \mathcal{W}^{-1}(Int(V_2(B))) & \textit{by the continuity and openness of } \mathcal{W} \\ & = & \mathcal{W}^{-1}(V_2(\Box B)) & \textit{by the definition of } V_2 \\ & = & \mathcal{W}^{-1}(V_2(A)). & \end{array}$$

- Case $A \equiv \bigcirc B$. We need to show that $V_1(\bigcirc B) = \mathcal{W}^{-1}(V_2(\bigcirc B))$. Let $x \in X_1$. We have

$$\begin{array}{rcll} x \in V_1(\bigcirc B) & \Leftrightarrow & T_1(x) \in V_1(B) & \textit{by the definition of } V_1 \\ & \Leftrightarrow & T_1(x) \in \mathcal{W}^{-1}(V_2(B)) & \textit{by the induction hypothesis} \\ & \Leftrightarrow & \mathcal{W}(T_1(x)) \in V_2(B) & \\ & \Leftrightarrow & T_2(\mathcal{W}(x)) \in V_2(B) & \textit{since } \mathcal{W} \textit{ is a functor} \\ & \Leftrightarrow & \mathcal{W}(x) \in V_2(\bigcirc B) & \textit{by the definition of } V_2 \\ & \Leftrightarrow & x \in \mathcal{W}^{-1}(V_2(\bigcirc B)). & \end{array}$$

QED

LEMMA 10.55 *Let $M_1 = \langle X_1, T_1, V_1 \rangle, M_2 = \langle X_2, T_2, V_2 \rangle$ be two dynamic topological models. Suppose that $\mathcal{W} : M_1 \to M_2$ is a commuting map and for each propositional variable p,*

$$V_1(p) = \mathcal{W}^{-1}(V_2(p)).$$

Then for any formula A of $L^{\bigcirc\Box}$,

$$M_2 \models A \text{ iff } M_1 \models A.$$

Proof Suppose that $M_2 \models A$, that is, $V_2(A) = X_2$. By Lemma 10.54 $V_1(A) = \mathcal{W}^{-1}(V_2(A))$, and so $V_1(A) = X_1$ as required. On the other hand suppose that $M_1 \models A$, but $M_2 \not\models A$, i.e., $V_2(A) \neq X_2$. Since $\mathcal{W}$ is onto and $V_1(A) = \mathcal{W}^{-1}(V_2(A))$, we have $V_1(A) \neq X_1$, that is, $M_1 \not\models A$, a contradiction. QED

COROLLARY 10.56 *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two classes of dynamic Kripke models such that for every model $M_2 \in \mathcal{C}_2$ there is an $M_1 \in \mathcal{C}_1$ and a functor $\mathcal{W}$: $M_1 \to M_2$. Then, if $\mathcal{C}_2$ is complete for S4C, then $\mathcal{C}_1$ is also complete for S4C.*

Proof If $M_2 \not\models A$, then $M_1 \not\models A$ by Lemma 10.55. QED

7.4 Stratified frames and limits

This section uses definitions and results of Slavnov, 2005 as reformulated in Fernandez, 2006.

DEFINITION 10.57 *Given a finite Kripke frame $W = \langle W, R\rangle$ we say that* limits are chosen in W *if for any R-monotone sequence $\{w_n\}$ a particular element $w = w_i$ is fixed such that Rw_nw for all n. We write $w = \lim_n\{w_n\}$. We assume that if $\{w_n\}$ stabilizes, that is $w_n = w$ for $n \geq n_0$, then $w = \lim_n\{w_n\}$.*

In Sec. 5 (and earlier in Mints, 1999) we chose limits in an arbitrary way. Now, given a dynamic system $\langle X, g\rangle$, we would like the function g to commute with the limits.

THEOREM 10.58 S4C *is complete for the class of finite dynamic Kripke models $\langle W, R, g, V\rangle$ where limits are chosen so that*

$$g(\lim_n\{w_n\}) = \lim_n\{g(w_n)\}.$$

Proof Consider a formula α refuted in a world v_0 of a finite dynamic Kripke model $M = \langle W, R, g, V\rangle$ where limits are chosen in some way. Let's "stratify" M. Let d be the $\bigcirc$-depth of α, that is the maximal nesting of $\bigcirc$ in α. Define

$$\tilde{W} = \{\langle w, g(w), \ldots, g^j(w)\rangle : w \in W,\ 0 \leq j \leq d\}$$

$$\tilde{R}\langle w, g(w), \ldots, g^j(w)\rangle\langle v, g(v), \ldots, g^k(v)\rangle \iff Rg^j(w)g^k(v) \ \&\ j = k$$

$$\langle w, g(w), \ldots, g^j(w)\rangle \in \tilde{V}(p) \iff g^j(w) \in V(p)$$

$$\tilde{g}(\langle w, g(w), \ldots, g^j(w)\rangle) := \langle w, g(w), \ldots, g^j(w), g^{j+1}(w)\rangle$$

for $j < d$;

$$\tilde{g}(\langle w, g(w), \ldots, g^d(w)\rangle) := \langle w, g(w), \ldots, g^d(w)\rangle$$

$$\lim_n \{\langle w_n, g(w_n), \ldots, g^j(w_n)\rangle\} :=$$

$$\{\langle \lim_n \{w_n\}, g(\lim_n \{w_n\}), \ldots, g^j(\lim_n \{w_n\})\rangle\}.$$

$\tilde{M} := \langle \tilde{W}, \tilde{R}, g, \tilde{V}\rangle$ is a dynamic topological model.

Consider an example. Let $M = \langle \{0, 1, 2\}, R, g, V\rangle$ with $R00$, $R01$, $R02$, $R11$, $R12$, $R21$ and $R22$; with $g(0) = g(2) = 1$ and $g(1) = 2$; and with any valuation function V. Note that 0 is the root and that $\{1, 2\}$ is a cluster. The original model M is shown at the left of Fig. 10.6. For $M = M_0$ and depth

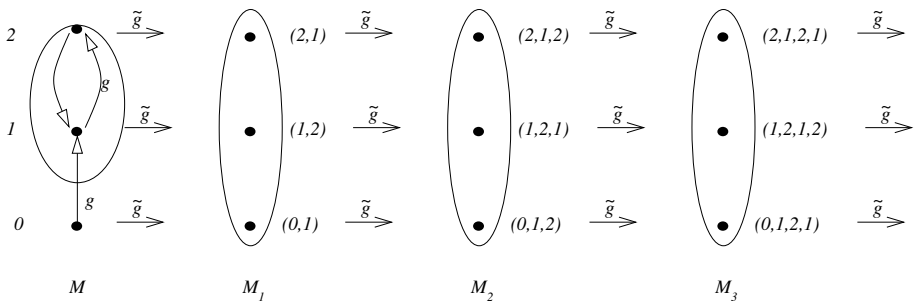


Figure 10.6. $\tilde{M} = M_0 \cup M_1 \cup M_2 \cup M_3$.

$d = 3$ we have $\tilde{M} = M_0 \cup M_1 \cup M_2 \cup M_3$, as shown in Fig. 6. Each of the M_i, $i > 0$, is a cluster, and the function $\tilde{g}$ moves M_i to M_{i+1} "horizontally".

In the general case, induction on formulas shows that

$$\langle w, g(w), \ldots, g^j(w)\rangle \in \tilde{V}(\phi) \iff g^j(w) \in V(\phi)$$

for all $j \leq d$ and all ϕ of the $\bigcirc$-depth $\leq d - j$. In particular $(w) \in \tilde{V}(\phi)$ iff $w \in V(\phi)$, for every ϕ and every $w \in W$, so given formula α is refuted at (v_0). Note that $\tilde{g}$ commutes with limits by the definition. QED

7.5 Completeness of S4C for Cantor Space

7.5.1 Setup. Here we present a streamlined version from Kremer, 2004 of a proof from Mints and Zhang, 2005b that S4C is sound and complete for Cantor Space. To state our main theorem, we define both Cantor Space and a number of related spaces based on trees.

Let ${}^*\omega$ be the set of finite sequences of natural numbers, including the empty sequence Λ. We use bold $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{v}$ to range over ${}^*\omega$. For $\mathbf{x}, \mathbf{y} \in {}^*\omega$, we write $\mathbf{xy}$ or $\mathbf{x}^\frown\mathbf{y}$ for $\mathbf{x}$ concatenated with $\mathbf{y}$. For $\mathbf{x}, \mathbf{y} \in {}^*\omega$ we say that $\mathbf{x} \leq \mathbf{y}$ iff $\mathbf{x}$ is an initial segment of $\mathbf{y}$, and $\mathbf{x} < \mathbf{y}$ iff $\mathbf{x} \leq \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$. For $\mathbf{x} \in {}^*\omega$ and $k \in \omega$, we write $\mathbf{x}k$ or $\mathbf{x}^\frown k$ for $\mathbf{x}$ concatenated with k. For $\mathbf{x} \in {}^*\omega$, $length(\mathbf{x})$ is the length of $\mathbf{x}$; and for $m < length(\mathbf{x})$, $\mathbf{x}_m$ is the mth member of $\mathbf{x}$. So $\mathbf{x} = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{length(\mathbf{x})-1}$.

For our purposes, a *tree* is any $T \subseteq {}^*\omega$ satisfying the following:

$$\Lambda \in T;$$

$$\forall\, \mathbf{x} \in T, \exists\, k \in \omega, \mathbf{x}k \in T;$$

and

$$\forall\, \mathbf{x} \in^* \omega, \forall\, k \in \omega, \text{ if } \mathbf{x}k \in T \text{ then } \mathbf{x} \in T.$$

We can think of the members of T as nodes in T. T is *nontrivially branching* iff $\forall\, \mathbf{x} \in T, \exists\, k, j \in \omega,\ k \neq j$ and $\mathbf{x}k, \mathbf{x}j \in T$. T is *finitely branching* iff $\forall\, \mathbf{x} \in T$, the set $\{k \in \omega : \mathbf{x}k \in T\}$ is finite.

We will be interested in *paths* through trees: these can be represented by members of ${}^\omega\omega$, i.e. by ω-long sequences of natural numbers. We use bold-italic $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{v}$ to range over ${}^\omega\omega$. For $\mathbf{x} \in {}^*\omega$ and $\boldsymbol{y} \in {}^\omega\omega$, we write $\mathbf{x}\boldsymbol{y}$ or $\mathbf{x}^\frown\boldsymbol{y}$ for $\mathbf{x}$ concatenated with $\boldsymbol{y}$. For $\boldsymbol{x} \in {}^\omega\omega$ and $m \in \omega$, $\boldsymbol{x}_m$ is the mth member of $\boldsymbol{x}$. So $\boldsymbol{x} = \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \ldots$. And for $\boldsymbol{x} \in {}^\omega\omega$ and $m \in \omega$, $\boldsymbol{x}{\restriction}m$ is the finite sequence consisting of the first m member of $\boldsymbol{x}$. So $\boldsymbol{x}{\restriction}m = \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m-1}$. If T is a tree, a *T-path* is an $\boldsymbol{x} \in {}^\omega\omega$ such that $\boldsymbol{x}{\restriction}m \in T$ for every $m \in \omega$.

For every tree T we define

$$path(T) = \{\boldsymbol{x} : \boldsymbol{x} \text{ is a } T\text{-path}\}.$$

A *T-path through the node* $\mathbf{x} \in T$ is any T-path of the form $\mathbf{x}^\frown\boldsymbol{y}$. We impose a *topology* on $path(T)$ as follows. For $\mathbf{x} \in T$, let

$$\begin{aligned} \mathbf{B}^T_{\mathbf{x}} &= \{\mathbf{x}^\frown\boldsymbol{y} : \boldsymbol{y} \in {}^\omega\omega \text{ and } \mathbf{x}^\frown\boldsymbol{y} \in path(T)\} \\ &= \text{the set of } T\text{-paths through the node } \mathbf{x}. \end{aligned}$$

Generalized T-Cantor Space is $path(T)$ with the topology determined by the basis sets $\mathbf{B}^T_{\mathbf{x}}$, where $\mathbf{x} \in T$. And *Cantor Space* is $\mathcal{C} = path({}^*2)$, where ^{*2} is the set of finite sequences of 0 and 1.

Our main theorem is as follows.

THEOREM 10.59 (SOUNDNESS AND COMPLETENESS IN CANTOR SPACE) $\mathcal{C} \models A$ *iff* $A \in$ S4C.

Since soundness (i.e. the $\Leftarrow$ direction of the " iff") is routine, we concentrate on completeness: if $\mathcal{C} \models A$ then $A \in$ S4C. Or, as we will prove it, if $A \notin$

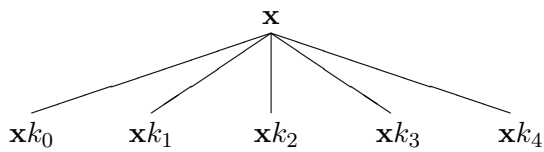
S4C then $\mathcal{C} \not\models A$. First, we reduce the problem of showing completeness in $\mathcal{C}$ to the problem of showing completeness in $path(T)$, where T is any finitely branching tree.

LEMMA 10.60 *Suppose that T is a finitely branching tree. Then, for every formula A, we have $\mathcal{C} \models A$ iff $path(T) \models A$.*

Proof First assume T is nontrivially branching and construct a homeomorhism of $path(T)$ onto $\mathcal{C}$, that is one-one onto continuous open function, $\phi : path(T) \to \mathcal{C}$. For this we construct a one-one function $\psi : T \to {}^*2$. Let

$$\psi(\Lambda) = \Lambda.$$

Suppose that $\psi(\mathbf{x})$ is defined for $\mathbf{x} \in T$, and let $k_0, \ldots, k_{m-1}$ be an exhaustive strictly increasing list of natural numbers such that $\mathbf{x}k_n \in T$ for each $n < m$. Note that m is finite since T is finitely branching, and that $m > 1$ since T is nontrivially branching. If $m = 5$ then we have the following picture of a portion of the tree T:

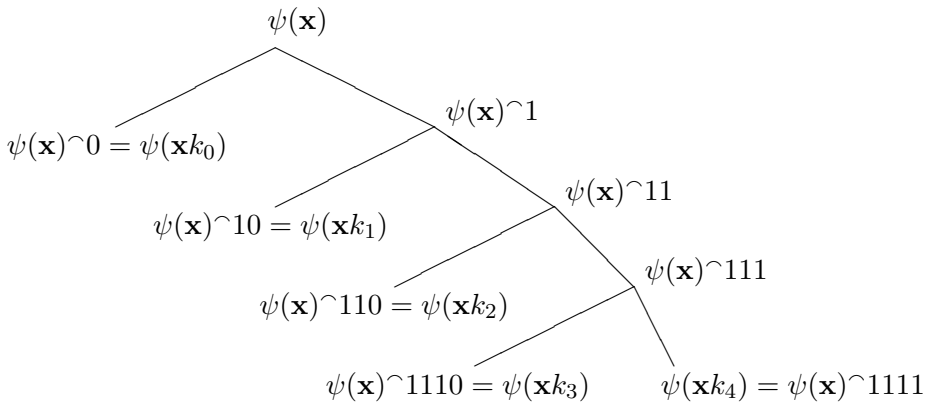


We define $\psi(\mathbf{x}k_n)$ as follows, for each $n < m$:

$$\psi(\mathbf{x}k_n) = \psi(\mathbf{x})^\frown 1^n{}^\frown 0 \text{ if } n < m-1,$$

$$\psi(\mathbf{x})^\frown 1^n \text{ if } n = m-1.$$

In this definition, 1^n is the finite sequence consisting of n occurrences of 1. The portion of T pictured above would get mapped by ψ into a portion of ^{*2} as follows:



The function $\psi : T \to^* 2$ induces a function $\phi : path(T) \to \mathcal{C}$ as follows: $\phi(\boldsymbol{x})$ is the unique ^{*2}-path through the nodes $\psi(\boldsymbol{x}\restriction 0), \psi(\boldsymbol{x}\restriction 1), \psi(\boldsymbol{x}\restriction 2), \psi(\boldsymbol{x}\restriction 3), \ldots$. Note that ϕ is indeed an isomorphism, hence T and $\mathcal{C}$ verify the same formulas.

Now consider case when some points $\mathbf{x} \in T$ have only one successor $\mathbf{x}k$. Construct a new tree T' by duplicating every such successor and the whole branch it begins.

$$T_1 := \{\mathbf{x}_1{}^\frown k_1' \ldots {}^\frown \mathbf{x}_l{}^\frown k_l' : \ \mathbf{x}_1{}^\frown k_1 \ldots {}^\frown \mathbf{x}_l{}^\frown k_l \in T \text{ and C}\}$$

where C means: each of $k_1, \ldots k_l$ is the only successor in T, $k_i' \in \{k_i, k_i + 1\}$ and at least one of k_i' is $k_i + 1$.

$$T' = T \cup T_1$$

T' is a nontrivially branching tree. Define a map $\mathcal{W} : T' \to T$ sending each new node into the old node of the same length from which it was generated:

$$\mathcal{W}(\mathbf{x}_1{}^\frown k_1' \ldots {}^\frown \mathbf{x}_l{}^\frown k_l') := \mathbf{x}_1{}^\frown k_1 \ldots {}^\frown \mathbf{x}_l{}^\frown k_l \in T \text{ for } \mathbf{x}_1{}^\frown k_1' \ldots {}^\frown \mathbf{x}_l{}^\frown k_l' \in T_1$$

$\mathcal{W}(\mathbf{x}) := \mathbf{x}$ for $\mathbf{x} \in T$. $\mathcal{W}$ is continuous, open and onto T. The inclusion map $I : T \to T'$ is " inverse" to $\mathcal{W}$. If $f : T \to T$ is a continuous function, it induces a continuous function $g : T' \to T'$:

$$g(\mathbf{x}) = f(\mathcal{W}(\mathbf{x}))$$

such that $\mathcal{W}$ commutes with f, g. Indeed,

$$\mathcal{W}(g(\mathbf{x})) = \mathcal{W}(f(\mathcal{W}(\mathbf{x})) = f(\mathcal{W}(\mathbf{x}))$$

since the values of f belong to T. Hence T and T' verify the same formulas by Lemma 10.55. QED

Given Lemma 10.60 and Theorem 10.52, proving Theorem 10.59, reduces to proving the following:

THEOREM 10.61 *If the formula ϕ is refuted by some rooted finite dynamic Kripke model, then $path(T) \not\models \phi$ for some finitely branching tree T.*

Proof So suppose that the formula ϕ is refuted by some finite dynamic Kripke model $M' = \langle W', R', f', V' \rangle$. Consider finite dynamic Kripke model

$$M := \tilde{M}' = \langle W, R, f, V \rangle$$

constructed from M' as in the proof of Theorem 10.58. The limits $\lim_n$ are chosen so that f commutes with limits. Let

$$\tilde{W} = \{0, 1, \ldots, m-1\}, \ m \geq 2.$$

Consider the tree of all R-monotone sequences of worlds $w \in W$ (that is of numbers $< m$). In other words, one-element sequences in T are

$$\langle 0\rangle, \langle 1\rangle, \ldots, \langle m-1\rangle$$

and sons (immediate successors) of an $\mathbf{x} = \langle x_0, \ldots x_n\rangle \in T$ are $x^\frown w_1, \ldots,$ $x^\frown w_l$, where $w_1, \ldots, w_l$ are all R-successors of x_n in W. This shows that for every $\boldsymbol{x} \in path(T)$ and all $n, l > 0$

$$R\boldsymbol{x}_n\boldsymbol{x}_{n+l}. \tag{10.18}$$

Define for $\boldsymbol{x} \in path(T)$

$$g(\boldsymbol{x}) := \lambda n.f(\boldsymbol{x}_n), \text{ that is } (g(\boldsymbol{x}))_n = f(\boldsymbol{x}_n)$$

and similarly for a finite sequence $\mathbf{x} \in T$

$$g(\mathbf{x}) = \lambda n.f(\mathbf{x}_n).$$

$g(\boldsymbol{x}), g(\mathbf{x})$ are R-monotone sequences by (10.18) and R-monotonicity of f, that is $g(\boldsymbol{x}) \in path(T)$ and $g(\mathbf{x}) \in T$. Moreover, g is continuous on $path(T)$ since $g(\boldsymbol{x})\restriction n$ is determined by $\mathbf{x} := \boldsymbol{x}\restriction n$, so that for $\mathbf{y} := g(\mathbf{x})$ we have for the corresponding basic open set $\mathbf{B}^T_{\mathbf{y}}$:

$$g^{-1}(\mathbf{B}^T_{\mathbf{y}}) \supseteq \mathbf{B}^T_{\mathbf{x}}.$$

Define for $\boldsymbol{x} \in path(T)$:

$$\mathcal{W}(x) := \lim_n\{x_n\}.$$

To finish, we prove that $\mathcal{W} : path(T) \to W$ is commuting (Definition 10.53).

$\mathcal{W}$ is continuous. Indeed, $\mathcal{W}(\boldsymbol{x}) = w$ implies $Rw\boldsymbol{x}, R\boldsymbol{x}w$ for all $n \geq n_0$. Hence for every $\boldsymbol{y}$ with $\boldsymbol{y}_{n_0} = \boldsymbol{x}_{n_0}$ and $n \geq n_0$ we have $Rw\boldsymbol{y}_n$. This implies $Rw \lim_n\{y_n\}$ that is $Rw\mathcal{W}(\boldsymbol{y})$. This proves $\mathcal{W}(y) \in O_w$, that is $\mathcal{W}^{-1}(O_w) \supseteq \mathbf{B}^T_{\boldsymbol{x}\restriction n_0}$.

$\mathcal{W}$ is open, since $w \in \mathcal{W}(\mathbf{B}^T_{\mathbf{y}})$ implies $O_w \subseteq \mathcal{W}(\mathbf{B}^T_{\mathbf{y}})$. Indeed, we have $\mathcal{W}(\boldsymbol{x}) = w$ for some $\boldsymbol{x}$ extending $\mathbf{y}$, implying $R\mathbf{y}_n w$ for $n = length(y) - 1$ (the last component of $\mathbf{y}$). For an arbitrary $w' \in O_w$ we have $R\mathbf{y}_n w'$. Define

$$\boldsymbol{z}_i := \begin{cases} \mathbf{y}_i & \text{if } i < n \\ w' & \text{if } i \geq n. \end{cases}$$

Then $\boldsymbol{z} \in \mathbf{B}^T_{\mathbf{y}}$ and $\mathcal{W}(\boldsymbol{z}) = \lim_i\{\boldsymbol{z}_i\} = w'$ as required. Finally $\mathcal{W}$ commutes with f, g, that is $g\mathcal{W} = f\mathcal{W}$:

$$\mathcal{W}(g(\boldsymbol{x})) = \lim_n\{g(\boldsymbol{x})_n\} = \lim_n\{f(\boldsymbol{x}_n)\} = f(\lim_n\{\boldsymbol{x}_n\}) = f(\boldsymbol{x}).$$

QED

REMARK 10.62 *We should note that, by an argument similar to the arguments in Sec. 6, we can show that*

$$\mathcal{C}, \mathcal{H} \models A \text{ iff } A \in S4\bigcirc,$$

where $\mathcal{C}$ is Cantor Space and $\mathcal{H}$ is the class of homeomorphisms. Thus S4$\bigcirc$ is the $\bigcirc\Box$ logic of homeomorphisms on $\mathcal{C}$ just as S4C is the $\bigcirc\Box$ logic of continuous functions of $\mathcal{C}$.

8. Conclusion

Let's outline the general picture and possible direction for future work. There are complete axiomatizations of general topology (S4) and of temporal logic (LTL). For the logic of dynamic systems, the "one-step" case in the language $L^{\bigcirc\Box}$ is also axiomatizable, and even decidable, both when the action $\bigcirc$ is only continuous and when it is a homeomorphism. Adding the trajectory connective $*p$ corresponding to $\&_n \bigcirc^n p$ results in an undecidable logic (Konev et al., 2006b). However significant fragments are (claimed to be) axiomatizable (Konev et al., 2006b) for continuous actions. On the other hand, the general case of the measure-preserving action–the case that provided the initial impetus for our investigation–turns out to be non-axiomatizable (Konev et al., 2006a). This does not prevent the possibility of a fruitful investigation by means of logic, similar to investigations in first order arithmetic or second order logic.

At this moment the most urgent task seems to be a propositional axiomatization allowing sufficiently many of the "routine" derivations in general topological dynamics, for example from (Aiken, 1993). Since the latter often involve a treatment of action as a relation, not just a function, a use of branching time temporal logic might be needed. Some of the problems left open by the current investigation may be of interest too, although it is difficult to predict which of these may be useful for mainstream mathematics. Let us mention the complete axiomatization of all continuous functions in the full trimodal language L (i.e. with modalities $\Box, \bigcirc$, and $*$) and an axiomatization in the language $L^{\bigcirc\Box}$ of continuous functions on the real segment, $[0, 1]$.

References

Aiello, M., van Benthem, J., and Bezhanishvili, G. (2003). Reasoning about Space: the Modal Way. *Journal of Logic and Computation*, 13(6):889–920.

Aiken, E. (1993). *The General Topology of Dynamical Systems*. American Mathematical Society.

Alexandrov, P. (1937). Diskrete Räume. *Matematicheskii Sbornik*, 2:501–518.

Artemov, S., Davoren, J., and Nerode, A. (1997). Modal Logics and Topological Semantics for Hybrid Systems. Technical Report MSI 97-05, Cornell University. Available at `http://web.cs.gc.cuny.edu/~sartemov/`.

Bezhanishvili, G. and Gehrke, M. (2005). A New Proof of Completeness of S4 with Respect to Real Line. *Annals of Pure and Applied Logic*, 133(1–3): 287–301.

Brown, J. (1976). *Ergodic Theory and Topological Dynamics*. Academic Press, New York.

Davoren, J. (1998). *Modal Logics for Continuous Dynamics*. PhD thesis, Cornell University.

Fernandez, D. (2006). *Completeness of S4C for* KMR^2.

Furstenberg, H. (1981). *Recurrence in Ergodic Theory and Combinatorial Number Theory*. Princeton University Press, Princeton.

Goldblatt, R. (1992). *Logics of Time and Computation*, volume 7 of *Center for the Study of Language and Information Lecture Notes*. Stanford University Press, Stanford, 2nd edition edition.

Konev, B., Kontchakov, R., Tishovsky, D., Wolter, F., and Zakharyaschev, M. (2006a). On Dynamic Topological and Metric Logics. *Studia Logica*. to be published.

Konev, B., Kontchakov, R., Wolter, F., and Zakharyaschev, M. (2006b). *Dynamic Topological Logics over Spaces with Continuous Functions*.

Kremer, P. (1997). Temporal Logic over S4: an Axiomatizable Fragment of Dynamic Topological Logic. *Bulletin of Symbolic Logic*, 3:375–376.

Kremer, P. (2004). *The Modal Logic of Continuous Functions on Cantor Space*. Availble at `http://individual.utoronto.ca/philipkremer/online papers/cantor.pdf`.

Kremer, P. and Mints, G. (1997). Dynamic Topological Logic. *Bulletin of Symbolic Logic*, 3:371–372.

Kremer, P., Mints, G., and Rybakov, V. (1997). Axiomatizing the Next-Interior Fragment of Dynamic Topological Logic. *Bulletin of Symbolic Logic*, 3: 376–377.

Kripke, S. (1963). Semantical Analysis of Modal Logic I, Normal Propositional Calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67–96.

McKinsey, J. C. C. and Tarski, A. (1944). The Algebra of Topology. *Annals of Mathematics*, 45:141–191.

Mints, G. (1999). A Completeness Proof for Propositional S4 in Cantor Space. In Orlowska, E., editor, *Logic at work. Essays dedicated to the memory of Helena Rasiowa.*, Stud. Fuzziness Soft Comput., chapter 24, pages 79–88. Heidelberg: Physica-Verlag.

Mints, G. and Zhang, T. (2005a). A Proof of Topological Completeness for S4 in (0,1). *Annals of Pure and Applied. Logic*, 133(1–3):231–246.

Mints, G. and Zhang, T. (2005b). Propositional Logic of Continuous Transformation in Cantor Space. *Archive for Mathematical Logic*, 44(6):783–799.

Rasiowa, H. and Sikorski, R. (1963). *The Mathematics of Metamathematics*. Państowowe Wydawnictwo Naukowe, Warsaw.

Segerberg, K. (1976). Discrete Linear Future Time Without Axioms. *Studia Logica*, 35:273–278.

Slavnov, S. (2003). Two Counterexamples in the Logic of Dynamic Topological Systems. Technical report, Cornell University.

Slavnov, S. A. (2005). On Completeness of Dynamical Topological Logic. *Moscow Mathematical Journal*, 5(5).

van Benthem, J. (1995). Temporal Logic. In Gabbay, D. M., Hogger, C. J., and Robinson, J. A., editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4, pages 241–350. Clarendon Press, Oxford.

Walters, P. (1982). *An Introduction to Ergodic Theory*. Springer-Verlag, Berlin.

Chapter 11

LOGIC OF SPACE-TIME AND RELATIVITY THEORY

Hajnal Andréka, Judit X. Madarász, István Németi
Rényi Mathematical Research Institute, Budapest

Second Reader

Valentin Shehtman
Moscow University & King's College London

1. Introduction

Our goal is to make relativity theory accessible and transparent for any reader with logical background. The reader does not have to "believe" anything. The emphasis is on the logic-based approach to relativity theory. The purpose is giving insights as opposed to mere recipes for calculations. Therefore proofs will be visual geometric ones, efforts will be made to replace computational proofs with suggestive drawings.

Relativity theory comes in (at least) two versions, special relativity (Einstein 1905) and general relativity (Einstein, Hilbert 1915). They differ in scope, the scope of general relativity is broader. Special relativity is a theory of motion and light propagation in vacuum far away from any gravitational object. I.e. special relativity does not deal with gravity. Also, special relativity is a "prelude" for general relativity, it provides a foundation or starting point for the general theory. General relativity unifies special relativity and the theory of gravitation. In some sense, general relativity is an "extension" of special relativity also putting gravity into the picture. General relativity can be used as a foundation for cosmology, e.g. it is a suitable framework for discussing the (evolution, properties of the) whole universe (expanding or otherwise). Special relativity, on the other hand, is not rich enough for this purpose. General relativity also provides the theory of black holes, wormholes, timewarps etc. Special relativity shows us that there is no such thing as space in itself, instead, a unified space-

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 607–711.
© 2007 *Springer.*

time exists. This inseparability of space and time becomes more dramatic in general relativity. Namely, general relativity shows us that gravity is nothing but the curvature of space-time. It is extremely difficult, if not impossible, to explain gravity without invoking the curvature (i.e. geometry) of space-time. The crucial point is that curvature of space is not enough (by far), it is space-time whose curvature explains gravity.[1] From a different angle: general relativity is a "geometrization" of much of what we know about the world surrounding us. E.g. it provides a full geometrization of our understanding of gravity and related phenomena like motion and light signals.

In Sec. 2 we study special relativity, in Sec. 3 we do the same for general relativity, in Sec. 4 we apply the so obtained tools to black holes, wormholes, timewarps. The emphasis is on the space-time aspects. Sec. 5 is about connections with the literature.

2. Special relativity

In this section, among others, we give a first-order logic (FOL) axiom system for special relativity such that we use only a handful of simple, streamlined axioms. In our approach, axiomatization is not the end of the story, but rather the beginning. Namely: axiomatizations of relativity are not ends in themselves (goals), instead, they are only tools. Our goals are to obtain simple, transparent, easy-to-communicate insights into the nature of relativity, to get a deeper understanding of relativity, to simplify it, to provide a foundation for it. Another aim is to make relativity theory accessible for many people (as fully as possible). Further, we intend to analyze the logical structure of the theory: which assumptions are responsible for which predictions; what happens if we weaken/fine-tune the assumptions, what we could have done differently. We seek insights, a deeper understanding. We could call this approach "reverse relativity" in analogy with "reverse mathematics".

2.1 Motivation for special relativistic kinematics in place of Newtonian kinematics

Why should we replace Newtonian Kinematics with such an exotic or counter-common-sense theory as special relativity? The Newtonian theory proved very successful for 200 years. By now, the Newtonian picture of motion has become the same as the current common-sense picture of motion. Hence the

[1] If we took into account the curvature of space only, then apples would no more fall down from trees. Gravitational attraction as such would disappear. On the other hand, if we keep the temporal aspects of curvature but ignore curvature of pure space, then gravity would not disappear, instead, this would cause only minor discrepancies in predicting trajectories of very fast moving bodies (relative to the source of gravity, e.g. the Earth or a black hole).

question is why we have to throw away our common-sense understanding of motion.[2]

The answer is that there are several independently good reasons for replacing the Newtonian worldview with relativity. These reasons are really good and decisive ones. They are so compelling, that any one of them would be sufficient for justifying and motivating our replacing the Newtonian worldview with relativity. We will mention a few of these reasons, but for simplicity of presentation, we will base this work on a fixed one of these reasons, namely on the outcome of the Michelson-Morley experiment. We will call this outcome of the Michelson-Morley experiment the Light Axiom. There are deeper, more philosophical reasons for replacing the Newtonian worldview with relativity theory, which might convince readers who are not experimentally minded, i.e. who are not easily convinced by mere facts about how results of certain experiments turned out. These philosophical reasons (under the name "principles of relativity") are intimately intertwined with issues which were significantly present through the last 2500 years of the history of our culture; see p. 663 and Barbour, 1989.

We now turn to the Light Axiom which will play a central role in this chapter. The first test of the Light Axiom was the Michelson-Morley experiment in 1887 and it has been tested extremely many times and in many radically different ways ever since. As a consequence, the Light Axiom has been generally accepted. An informal, intuitive formulation of the axiom is the following. (Later we will present this axiom in more formal, more precise terms, too, see AxPh in Sec. 2.3.)

Light Axiom: The speed of light is finite and direction independent, in the worldview of any inertial observer.

In other words, the Light Axiom means the following. Imagine a (huge) spaceship drifting through outer space in inertial motion. (*Inertial* here means that the rockets of the spaceship are switched off, and that the spaceship is not spinning.) Assume a scientist in this inertial spaceship is making experiments. The claim is that if the scientist measures the speed of light, he will find that this speed is the same in all directions and that it is finite. It is essential here that this is claimed to hold for all possible inertial spaceships irrespective of their velocities relative to the Earth or the Sun or the center of our galaxy or whatever reference system would be chosen. The point is that no matter which inertial spaceship we choose, the speed of light in that spaceship is independent of the direction in which it was measured, i.e. it is "isotropic".

In the technical language what we called "inertial spaceship" above is called an inertial reference frame, and the scientist in the spaceship making the exper-

[2] A second, equally justified question would ask why exactly those postulates/axioms are assumed in relativity which we will assume. We will deal with both questions.

iments is called an "observer". Later "*observer*" and reference frame tend to be identified.[3]

Let us notice that the Light Axiom is surprising, it is in sharp contrast with common-sense. Namely, common-sense says that if we send out a light signal from Earth, and a spaceship is racing with this light signal moving with almost the speed of the signal in the same direction as the signal does, then the velocity of the signal relative to the spaceship should be very small. Hence, one would think that the astronaut in the spaceship will observe the motion of the light signal as very slow. With the same kind of reasoning, the astronaut should observe light signals moving in the opposite direction very fast. But the Light Axiom states that light moves with the same speed in all directions for the astronaut in the spaceship, too. Hence, the Light Axiom flies in the face of common-sense. This gives us a hint/promise that very interesting, surprising things might be in the making. See also Fig. 11.18 on p. 642.

In fact, if we add the Light Axiom to Newtonian Kinematics, then we obtain a logical contradiction. I.e. (Newtonian Kinematics + Light Axiom) is an inconsistent theory in the usual sense of logic as we will outline soon (cf. Proposition 11.1). Seeing this contradiction, Einstein did the natural thing. He weakened Newtonian Kinematics (NK for short) to a weaker theory NK^- such that NK^- became consistent with the Light Axiom. Then the theory (NK^- + Light Axiom) became known as Special Relativistic Kinematics (SRK for short). We will study this theory under the name $\mathbf{Specrel}_0$ to be introduced in a logical language soon. We represent the above outlined process by the following diagram:

(NK + Light Axiom) **leads to Contradiction** (!)

$$\Downarrow$$

NK **gets replaced with the weaker** NK^-

$$\Downarrow$$

(NK^- + Light Axiom) **receives the name** Special Relativity (SRK).

SRK **is consistent** (this will be proved in Corollary 11.12, p. 644).

To see the above process more clearly, let us invoke a possible axiomatization of NK, still on the intuitive level.

Preparation for NK: If we want to represent motion of "particles" or "bodies" or "mass-points", it is natural to use a 4-dimensional Cartesian coordinate system $\mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (where $\mathbb{R}$ is the set of real numbers), with one time dimension t and three space dimensions x, y, z. A three-dimensional part of

[3]However, it is good to keep in mind that some thought-experiments are carried out by a team of observers (and if the members of this team do not move relative to each other then they are called, for simplicity, a single observer).

this is depicted in Fig. 11.1. The time-axis t is drawn vertically. Representing

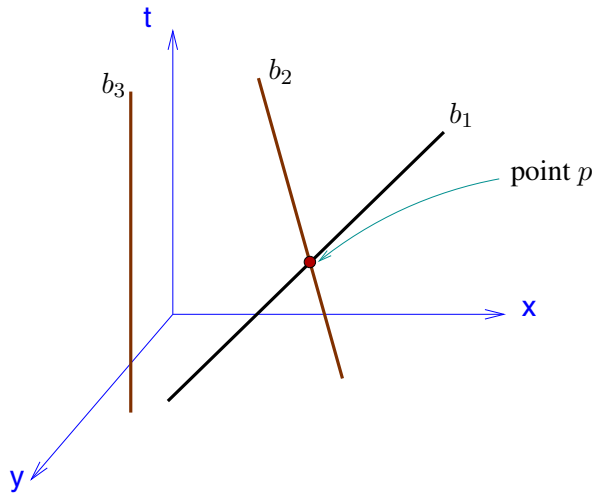


Figure 11.1. A space-time diagram. Wordlines of bodies b_1, b_2, b_3 represent motion. (Coordinate z is not indicated in the figure.) b_3 is motionless and b_1 moves faster than b_2.

the motion of a body, say b, in a 4-coordinate system can be done by specifying a function f which to each time instance $t \in \mathbb{R}$ tells us the space coordinates x, y, z where the body b is found at time t. Hence a function f : Time $\rightarrow$ Space specifies motion of a particle in this sense. The function f representing motion of b is called the *worldline* or lifeline of b. Fig. 11.1 represents motion of bodies, in this spirit. Besides the coordinate axes, we have represented the worldlines of inertial bodies b_1, b_2 and b_3 in Fig. 11.1. The straight line labeled by b_1 is the worldline of b_1. The slope of the worldline of b_1 is greater than that of b_2 which means that b_1 moves faster than b_2 does. The worldline of the third body b_3 is parallel with the time-axis, this means that b_3 is motionless. Bodies b_1 and b_2 meet at space-time point $p = \langle t, x, y, z \rangle$. Such a meeting (of two or more bodies) is called an *event*. We will extensively refer to such 4-dimensional coordinate systems and such worldlines of bodies and events.

The axioms of NK are summarized as (i)-(v) below.

(i) Each observer "lives" in a 4-coordinate system as described above. The observer in his own coordinate system is motionless in the origin, i.e. his worldline is the time-axis.[4]

(ii) Inertial motion is straight: Let o be an arbitrary inertial observer and let b be an inertial body. Then in o's 4-coordinate system the worldline of b is a

[4]It is sufficient to assume that his worldline is parallel with the time-axis.

straight line. I.e. in an inertial observer's worldview or 4-coordinate system all worldlines of inertial bodies appear as straight lines.

As we said, an observer in his metaphorical "spaceship" is inertial if his rockets are turned off and the spaceship is not spinning. In special relativity, we discuss only inertial motion, hence in our axiomatization the adjective "inertial" could be omitted. (Of course, then we need a general claim that only inertial things/objects will be studied.)

(iii) Motion is permitted: In the worldview or 4-coordinate system of any inertial observer it is possible to move through any point p in any direction with any finite speed.

(iv) Any two observers "observe" the same events. I.e. if according to o_1 bodies b_1 and b_2 have met, then the same is true in the 4-coordinate system of any o_2. We postulate the same for triple meetings e.g. of b_1, b_2, b_3.

(v) Absolute time: Any two observers agree about the amount of time elapsed between two events. (Hence temporal relationships are absolute.)

So, now, NK is defined as the theory axiomatized by (i)-(v) above.

It is easy to see that (NK + Light Axiom) is inconsistent. Einstein's idea was to check which ones of (i)–(v) are responsible for contradicting the Light Axiom and to throw away or weaken the "guilty" axioms of NK. We will see that (v) is guilty and that part of (iii) is suspicious. Hence we throw away (v) and weaken (iii) to a safer form (iii$^-$) where (iii$^-$) is the following.

(iii$^-$) Slower-than-light motion is possible: in the worldview of any inertial observer, through any point in any direction it is possible to move with any speed slower than that of light (here, light-speed is understood as measured at that place and in that direction where we want to move).

In the formal part we will carefully study whether all of these modifications are really needed and to what extent (cf. Theorems 11.4, 11.7). We define NK$^-$ as the remaining theory:

$$\mathrm{NK}^- := \{(\mathsf{i}), (\mathsf{ii}), (\mathsf{iii}^-), (\mathsf{iv})\}$$

and *Special Relativistic Kinematics* is defined as

$$\mathrm{SRK} := (\mathrm{NK}^- + \textsf{Light Axiom}).$$

The formalized version of this SRK will appear later as the theory $\mathbf{Specrel}_0$. We will prove that $\mathbf{Specrel}_0$ is consistent (i.e. contradiction-free) and will study its properties. Therefore SRK is also consistent, since, as we said, $\mathbf{Specrel}_0$ is a formalized version of SRK. Actually, the whole process of arriving from NK and the Light Axiom (or some alternative for the latter) to SRK will be subjected to logic-based conceptual analysis in Sec. 2.5.

Before turning to formalizing (and studying) Special Relativity SRK in logic, let us prove (informally only) on the present level of precision why absolute

time (i.e. axiom (v)) is excluded by the Light Axiom, or more precisely, it is excluded if we want to keep a fragment of our intuitive picture of the world, i.e. if we want to keep (i), (ii), (iv) of NK. We will prove:

$$(\mathrm{NK}^- + \text{Light Axiom}) \vdash \text{Negation of (v)},$$

where we use turnstile "$\vdash$" as the symbol of logical provability or derivability. I.e. $A \vdash B$ means that from statement A one can prove, rigorously, statement B.

Actually, we will prove something stronger and stranger from the Light Axiom (and a fragment of NK^-). We will prove that the time elapsed between two events may be different for different observers even in the special case when this elapsed time is zero for one of the observers. I.e. the very question whether two events happened at the same time or not will depend on the observer: two events A and B may happen at the same time for me, while event A happened much later than event B for the Martian in his spaceship. We will refer to this phenomenon by saying that "*simultaneity is not absolute*". Moreover, we will see later (Corollary 11.5) that the temporal order of some events may be switched: event A may precede event B for me, while for the Martian in his spaceship, event B precedes event A.

We say that events e and e' are *simultaneous* for observer O if in O's coordinate system the two events e, e' happen at the same time.

PROPOSITION 11.1 (SIMULTANEITY IS NOT ABSOLUTE) *Assume* SRK. *Moving clocks get out of synchronism, i.e.: Assume that a spaceship S is in uniform motion relative to another one, say E, and assume that two events e, e' happen simultaneously at the rear and at the nose of the spaceship S according to the spaceship S. Then e and e' take place at different times in E's coordinate system.*

I.e., the time elapsed between e and e' is zero as "seen" from the spaceship S, but the time elapsed between e and e' is nonzero as "seen" from E. See Fig. 11.3.

Intuitive proof Assume that we are in spaceship E, and let us call E "Earth". Assume that spaceship S—let us call it "Spaceship"—moves away from us in a uniform motion with, say, 0.9 light-speed. The captain of Spaceship positions his brothers called Rear, Middle, and Nose at the rear, middle and nose of the spaceship, respectively, and asks Rear and Nose to switch on their flashlights towards Middle exactly at the same time. Then the light signals (photons[5]) Ph1 and Ph2 from the two flashlights arrive to Middle at the same time, because

[5]We use the word "photon" as a synonym for light signal. It tacitly refers to the corpuscular conception of light. In this work we do not need the quantum-mechanical definition of photons. (That will be needed only in the final, as yet nonexistent, generalization of general relativity called quantum gravity.)

Middle is exactly in the middle of the spaceship, and because the speed of Ph1 sent by Rear is the same as the speed of Ph2 sent by Nose (by the Light Axiom). See Fig. 11.2.

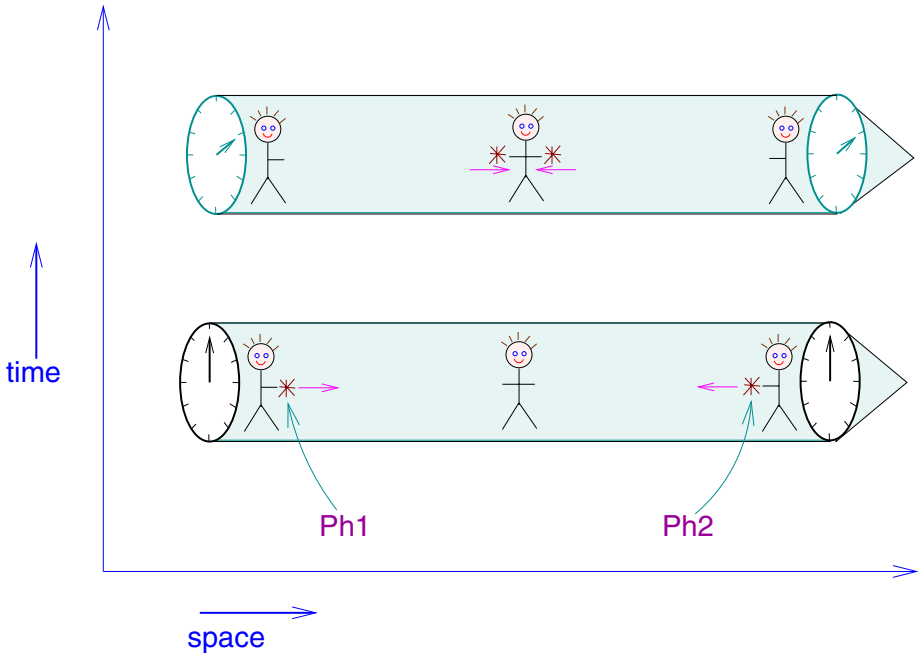


Figure 11.2. Seen from Spaceship, the two light-signals (i.e. photons) Ph1 and Ph2 are sent out at the same time, and meet in the middle. This is indicated by the clocks at the rear and at the nose of the spaceship. Notice that time in this figure is running upwards! I.e., this figure is similar to drawings in cartoons in that a sequence of scenes is represented in it. However, here the temporal order of the scenes is switched: the scene at the bottom took place earliest. The reason for this convention is our seeking compatibility with the usual space-time diagrams like Fig. 11.1.

How do we see all this from the Earth? We see that Rear and Nose send light signals (or photons) Ph1 and Ph2 towards Middle, and we also see that Ph1, Ph2 arrive to Middle at the same time (because this is a 3-meeting of bodies/entities and axiom (iv) in NK^-). (Spaceship's hull is missing, we can imagine it having only a grid of metal rods for keeping it together or something to this effect.) However, by the Light Axiom, the speeds of Ph1 and Ph2 are the same for us on the Earth, too. Since Spaceship moves away from us (with 0.9 light-speed), we see Ph1 crawl very slowly along the hull of Spaceship because the ship is "running away" from us (and from Ph1, too). On the other hand, the other photon, Ph2, flashes along the hull of the spaceship towards us with enormous

relative speed (relative to the hull of the spaceship). Because of this difference of their speeds relative to Spaceship, according to Earth, Ph1 and Ph2 either meet close to the rear of the spaceship, or if they meet in the middle, then Nose had to switch on his flashlight much later than Rear did. See Fig. 11.3.

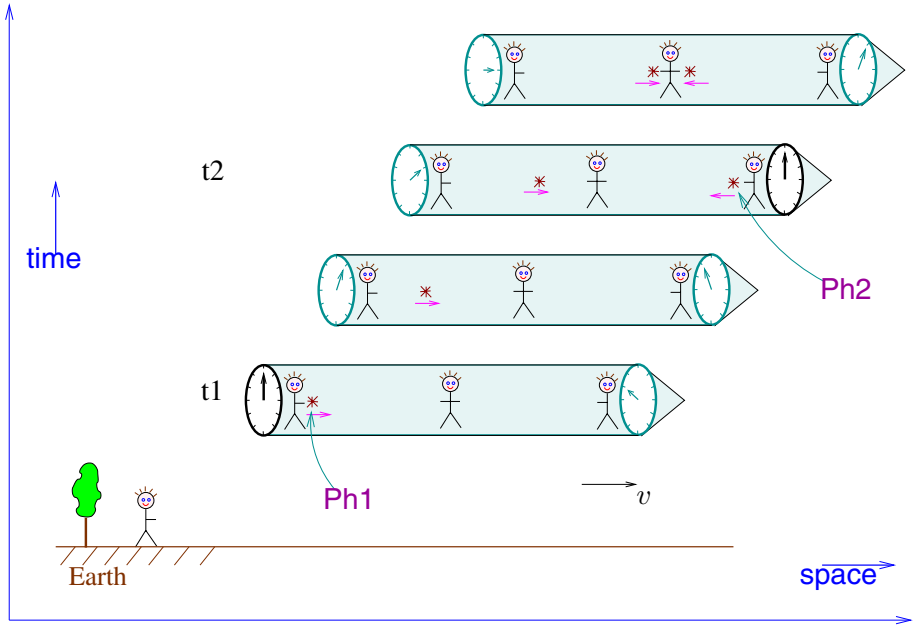


Figure 11.3. Seen from the Earth, the photon Ph2 had to be sent out later in order that it arrive in the middle at the same time as Ph1 does. But seen from Spaceship, they were sent out at the same time. Hence the clocks at the nose and the rear are out of synchronism, as seen from the Earth.

So far we proved that at least one of two things cannot be absolute. These are (a) being in the middle of the spaceship, and (b) simultaneity. Here, (a) means that Spaceship observes Middle in the middle of the ship, while Earth observes that Middle is not in the middle of the ship; and (b) means that emissions of photons Ph1 and Ph2 are simultaneous for Spaceship but not for Earth.

The first possibility is that Middle stands closer to the rear of Spaceship as seen from the Earth, i.e. that he is not in the middle of the ship according to Earth observers, while he is in the middle according to the ship observers. Here is a thought-experiment which shows that this is not possible. Let us ask the captain to give mirrors to Rear and Nose, and order Middle to send photons Ph3, Ph4 at the same time to these two mirrors. Since Middle is exactly in the middle of the ship, the bounced-back photons arrive to him at the same time, by the Light

Axiom. By (iv) in NK^-, we on the Earth also see that the two photons Ph3, Ph4 meet again at Middle after bouncing back, so they traveled their round-trips in the same amount of time. We will show that, as seen from the Earth, the time needed for the round-trip is proportional to the covered distance: if, say, Nose is twice as far from Middle as Rear is, then the time needed for Ph4 for the round-trip Middle-Nose-Middle is twice as much as the time needed for Ph3 for the round-trip Middle-Rear-Middle, even in a fast-moving spaceship. From the Earth we see that the round-trip took the same time for Ph3 and for Ph4, therefore we have to infer that Middle is really in the middle of the ship. See Fig. 11.4.

We now prove that the time needed for the round-trip is proportional to the covered distance. Indeed, assume that the distance Middle-Nose is twice as much as the distance Middle-Rear. We will show that the round-trip Middle-Nose-Middle takes twice as much time for a photon Ph4 as the round-trip Middle-Rear-Middle for a photon Ph3. Let us watch from the Earth how the two photons Ph3 and Ph4 move relative to the spaceship (as in Fig. 11.4, but now Middle standing closer to Rear). We will see that Ph3 covers the segment Middle-Rear fast, traveling towards us, and then covers the segment Rear-Middle slowly, moving away from us. The same way, Ph4 covers the segment Middle-Nose slowly, moving away from us, while Ph4 covers the segment Nose-Middle fast, moving towards us. The relative speed of Ph4 in the "towards-us" segment Nose-Middle is the same as the relative speed of Ph3 in the "towards-us" segment Middle-Rear; hence this part of the trip takes twice as much time for Ph4 as for Ph3 because we assumed that the distance Nose-Middle is twice as much as the distance Middle-Rear. The situation is completely analogous for the "away-from-us" segments, so the trip Middle-Nose takes twice as much time for Ph4 as the trip Rear-Middle for Ph3. Summarizing the segments, the round-trip takes twice as much time for Ph4 as for Ph3.

As we said earlier, we observe from the Earth that Ph3, Ph4 and Middle meet in a single event. Therefore, since we observe that Ph3 arrives to Middle exactly when Ph4 arrives to Middle after their round-trips, we have to infer, on the Earth, that Middle really stands exactly in the middle of Spaceship. There remains only the possibility that Nose sent out his photon Ph2, which we see as fast-moving along the hull of the space ship, much later than Rear sent Ph1 which we see as slowly-moving along the hull of the spaceship. Thus, as seen from the Earth, the clocks at the nose and at the rear of the spaceship show different times (at the same Earth-moment). This is what we mean when we say that clocks of the spaceship get out of synchronism.

Summing up: Let e and e' be the events when Rear sends his photon Ph1 towards Middle, and when Nose sends his photon Ph2 towards Middle, respectively. Then these two events took place at the same time as seen from

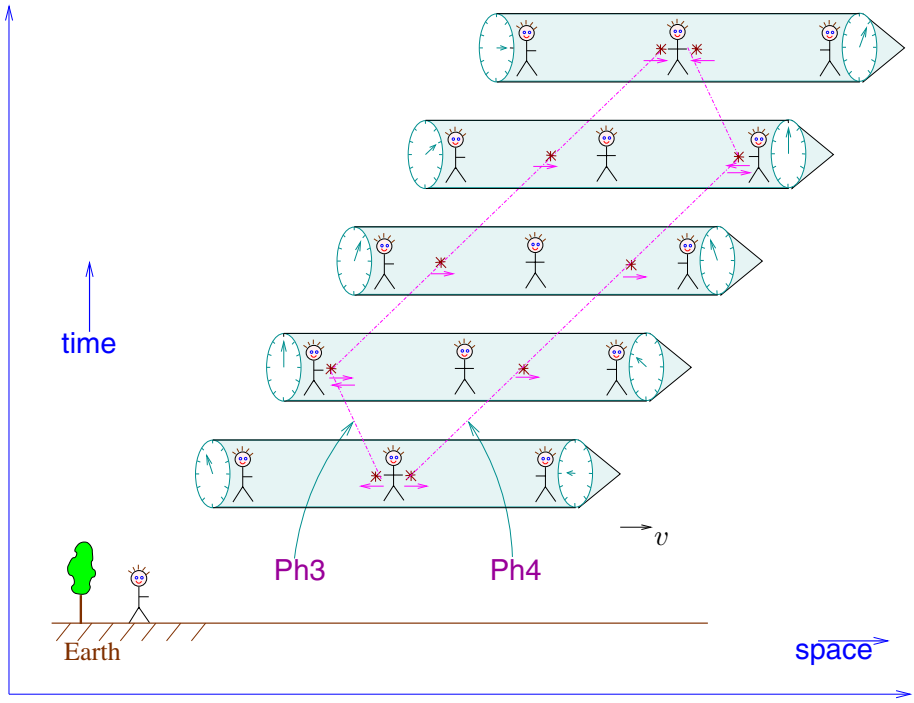


Figure 11.4. The round-trip for Ph3 takes the same time as for Ph4, seen both from Spaceship and from the Earth. Hence Earth infers that Middle is indeed in the middle of the ship.

Spaceship, while as seen from the Earth, e' took place later than e did. This finishes the proof of Proposition 11.1. QED

Let us notice that Proposition 11.1 above is a far reaching claim. It implies that one of the most basic words of natural language refers to an illusion only and carries no real meaning. The word in question is the word "now".

In order to be able to carry out the proof of Proposition 11.1 and other similar chains of thought in the "safe", precise setting of mathematical logic, in the next subsection we introduce a first-order language in which we formalize our axioms, statements, and proofs.

2.2 Language

Motivation for language. We want to talk about space-time as relativity theory conceives it. We will talk about space-time as experienced through motion. Though we discuss here kinematics (theory of motion) only, one can derive (logically) dynamic predictions of relativity, too (i.e. phenomena involving forces,

energy) using the same approach/axioms.[6] We will use first-order logic, and the most important decision is to choose the language (*vocabulary*), i.e. what objects and what relations between them will belong to the language. We will specify our first-order language by specifying its models.

We want to axiomatize motion. What moves? Bodies. Hence our model has a universe B for bodies. What does it mean to move? To move means changing location in time. We will have coordinate systems, or reference frames in other words, for marking locations and time, and we will use quantities in setting up these coordinate systems. So our model has another universe Q for quantities. We will think of quantities as real numbers, so Q together with the operations $+, *, <$ will form a linearly ordered field. We will think of coordinate systems as belonging to special bodies called observers. Hence Ob is a one-place relation on B (picking out a subset of B). We will have another kind of special bodies, photons, too. Hence Ph is another one-place relation on B. The heart of our model is the so-called worldview-relation W. This is a 6-place relation connecting bodies and quantities. We think of $W(o, b, t, x, y, z)$ as the statement that the body b is in location xyz at time t in observer o's coordinate system. We will simply pronounce this as

o sees the body b at $txyz$

though this has no connection with optical seeing, instead, it is an act of co-ordinatizing only. With this intuition in mind we now fix the language of our theories of special relativity.

The language. We fix a natural number $n > 1$, it will be the number of space-time dimensions. In most works $n = 4$, i.e. one has 3 space-dimensions and one time-dimension. Recent generalizations of general relativity in the literature indicate that it might be useful to leave n as a variable (e.g. string-theory uses 11 dimensions).

We declare two sorts of objects. One sort is for "quantities", it will be denoted by Q. (This is the same as "real numbers" in other treatments.) We have two-place (i.e. binary) operation symbols $+, *$ and a 2-place predicate symbol $<$ of sort "quantities". To avoid misunderstandings, we emphasize that, in this chapter, Q is not the set of rational numbers. (The letter Q abbreviates "quantities". It is a coincidence that the same letter is used in the literature to denote the rationals. We do not follow that convention.)

The other sort, B, is for entities which do the "moving". We will call these "*bodies*". (We call the moving entities "bodies" whatever they may be, in reality they can be e.g. coordinate systems or electromagnetic waves, or centers

[6]For the spirit of this we refer to the relativity textbook Rindler, 2001, Sec. 6 "Relativistic particle mechanics". Rindler, 2001, Sec. 6.2 is particularly relevant here.

of mass.) We have two kinds of special bodies, observers and photons. Thus Ob and Ph are one-place predicate symbols of sort B.

We have a relation which connects these two sorts, the $(n+2)$-place relation W which is of sort $\mathsf{B} \times \mathsf{B} \times \mathsf{Q} \times \cdots \times \mathsf{Q}$. The sentence "observer o observes body b at space-time location $p_1, \ldots, p_n$" is denoted as $\mathsf{W}(o, b, p_1, \ldots, p_n)$, or as $\mathsf{W}(o, b, p)$ in short.

Summing up, a model $\mathfrak{M}$ of our language is of form

$$\langle Q, +, *, <;\ \ B, Ob, Ph;\ \ W\rangle$$

where $\langle Q, +, *, <\rangle$ is a structure similar to ordered fields, Ob, Ph are subsets of B, and $W \subseteq B \times B \times Q \times \cdots \times Q$.

2.3 Axiomatization Specrel of special relativity in first-order logic

In this subsection we formalize the axioms we talked about on an intuitive level in Sec. 2.1, by using the first-order language introduced in the previous subsection. The formalized version of (NK$^-$ + Light Axiom) will be called **Specrel**$_0$.

Axiom 1 (AxField). The quantities behave like real numbers do in the sense that $\langle Q, +, *, <\rangle$ is a linearly ordered field in which every positive member has a square root. Such fields are called "quadratic".

For an axiom system for linearly ordered fields we refer to e.g. Chang and Keisler, 1973, p. 41, below item 18. We will often simply say "field" or "ofield" or "ordered field" instead of "linearly ordered field". We recall that if $\langle Q, +, *, <\rangle$ is a field, then 0 and 1 denote the neutral elements of $+$ and $*$ respectively, and an element $x \in Q$ is called *positive* iff $x > 0$. Further, y is called a *square root* of x iff $y * y = x$ and $y \geq 0$, we denote this by writing $y = \sqrt{x}$. With this notation, the *absolute value* $|y|$ of y is $|y| := \sqrt{y^2}$ (as usual, y^2 denotes $y * y$). The inverses of the operations $+$ and $*$ will be denoted by - and $/$, respectively. Thus a linearly ordered field is *quadratic* (or Euclidean) iff $(\forall x > 0)(\exists y)x = y * y$ is true in it. According to the usual practice, we will often omit $*$ from an expression, e.g. we write td in place of $t * d$.

On AxField: In most physics books, the set of quantities is taken to be the set of real numbers (with $+, *$ as addition, multiplication of the real numbers). Some, fancy, books use also imaginary numbers for quantities. We know from mathematics that much complexity is tied to the real numbers. Hence in our axiomatic approach, we single out those properties of the quantities that we rely on in the investigation in question. In special relativity only the quadratic ordered field-structure of the quantities is presupposed, but we could do much even with assuming only the ring-structure. In particular, the ordering and the existence of square roots are used mostly in order to be able to formulate

results in a simpler way. (E.g. we use square roots in expressing the distance between two coordinate points. Without the use of square roots we always would have to talk about the square of distance. This would cause only an inconvenience but not an impossibility.) As a pay-off of this explicit way of handling the quantities, we can build models of special relativity with a finite field as structure of quantities, or we can use fields with infinitesimally small numbers. In general relativity, in addition to AxField it will suffice to use an axiom-schema called CONT, see Sec. 3.6.

By a *coordinate point*, or *space-time location*, we understand an n-*tuple* (i.e. a sequence of length n) $p = \langle p_1, \ldots, p_n \rangle$ of elements of Q, the set of all these n-tuples is denoted by Q^n. If $p \in \mathsf{Q}^n$, then $p_1, \ldots, p_n$ are its components, i.e. $p = \langle p_1, \ldots, p_n \rangle$. We call $\overline{0} := \langle 0, \ldots, 0 \rangle \in \mathsf{Q}^n$ the (n-dimensional) *origin*, and we call $\overline{t} := \{\langle x, 0, \ldots, 0 \rangle \in \mathsf{Q}^n : x \in \mathsf{Q}\}$ the (n-dimensional) *time-axis*. By the *worldline* (or lifeline, or history) of a body b as observed by the observer m we mean the set of space-time locations where m observes b to be present,

$$\mathsf{wline}_m(b) := \{p \in \mathsf{Q}^n : \mathsf{W}(m, b, p)\}.$$

Axiom 2 (AxSelf) An observer m in his own coordinate system is motionless in the origin (of space), i.e. his worldline is the time-axis: $\mathsf{wline}_m(m) = \overline{t}$. As a formula of the FOL language this axiom is

$$(\forall m \in \mathsf{Ob})(\forall p \in \mathsf{Q}^n)[\,\mathsf{W}(m, m, p) \leftrightarrow p_2 = \cdots = p_n = 0\,].$$

Having a field in our language makes it possible to talk about straight lines. We recall that the *straight line* going through $p, q \in \mathsf{Q}^n, p \neq q$ is the set $\{p + x * (p - q) : x \in Q\}$. In the latter formula, $+, -$ and $*$ denote operations of Q^n as a *vector-space*. We will often say just "line" for "straight line".

Axiom 3 (AxLine) The motion of an observer as observed by any observer is uniform, i.e. such that both the "spatial direction" and the "pace" of the motion are constant (and "longest possible" with this property). In geometrical terms this means that in each observer's coordinate system, the worldline of an observer is a straight line, i.e. $\mathsf{wline}_m(k)$ is a straight line for all $m, k \in \mathsf{Ob}$. Formally,

$$\begin{aligned}&(\forall m, k \in \mathsf{Ob})(\exists p, q \in \mathsf{Q}^n)(\mathsf{W}(m, k, p) \wedge \mathsf{W}(m, k, q) \wedge p \neq q \wedge \\ &\quad (\forall r \in \mathsf{Q}^n)[\mathsf{W}(m, k, r) \;\leftrightarrow\; (\exists x \in \mathsf{Q}) r = p + x * (p - q)]).\end{aligned}$$

On AxLine: AxLine is a formalized version of postulate (ii) in Sec. 2.1. Later we will consider non-uniform motions, too. We will call those motions "*accelerated*" ones. Newton's First Law of Motion states that "an object moves with constant, uniform motion until acted on by a force". A body is called "inertial" if no force acts on it. Hence AxLine indicates that Ob denotes the set of inertial observers when using AxLine.

We introduce the *speed* of a uniform motion. In geometric terms, this is the "slant" or "*slope*" of the straight line (representing the motion). For $p, q \in \mathsf{Q}^n$, let $\mathsf{space}(p, q)$ and $\mathsf{time}(p, q)$ denote the *spatial distance* and the *time-distance* between p and q, respectively:

$$\mathsf{space}(p, q) := \sqrt{(p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}, \quad \text{and}$$

$$\mathsf{time}(p, q) := |p_1 - q_1| = \sqrt{(p_1 - q_1)^2}.$$

Now $\mathsf{speed}(p, q)$ denotes the speed necessary to reach q from p (or p from q):

$$\mathsf{speed}(p, q) := \mathsf{space}(p, q)/\mathsf{time}(p, q) \quad \text{when } \mathsf{time}(p, q) \neq 0.$$

Axiom 4 (AxPh) For every observer, the speed of light is 1, and moreover, photons move uniformly along straight lines and in each location in each direction it is possible to send out a photon. In geometrical terms this means that the worldlines of photons are exactly the straight lines of slope 1. Formally this is:

$$(\forall m \in \mathsf{Ob})(\forall \mathsf{ph} \in \mathsf{Ph})[\,(\mathsf{wline}_m(\mathsf{ph}) \text{ is a straight line}) \wedge (\forall p, q \in \mathsf{Q}^n)$$
$$p \neq q \Rightarrow (\mathsf{speed}(p, q){=}1 \text{ iff } (\exists \mathsf{ph} \in \mathsf{Ph})[\mathsf{W}(m, \mathsf{ph}, p) \wedge \mathsf{W}(m, \mathsf{ph}, q)])].$$

On AxPh: This is the formal version of the Light Axiom used in Sec. 2.1. It expresses that the speed of light is finite (nonzero) and *isotropic*, i.e. direction-independent. We formulated the Light Axiom in a seemingly stronger form, namely such that we require the speed of light to be 1. This way we are freed from having to deal with always adjusting everything to the actual speed of light. Instead, we adjust the units of measurement to the speed of light: we measure distances with "light-years" if we measure time in "years". We emphasize that assuming that the speed of light is 1 instead of some finite direction independent number (which might depend on the observer) is not a "physical" assumption but instead a merely "linguistic" one. It would be sufficient (for our results) to use a more literal formalization of the Light Axiom in Sec. 2.1. That such a weaker version of AxPh is sufficient for our results in shown in Andréka et al., 2002 and Madarász, 2002, p. 121.

In Sec. 2.1 we talked about photons bounced back from a mirror. When using AxPh, we will simulate this "bouncing back" by treating the out-going and the bounced-back photons as two different photons that have met at the mirror (see e.g. Fig. 11.8).

The next axiom states that each observer can make thought-experiments in which he assumes the existence of "slowly moving" observers. This is the formalized version of postulate (iii$^-$) in Sec. 2.1.

Axiom 5 (AxThEx) For each observer $m \in \mathsf{Ob}$, in each space-time location, in each direction, with any speed smaller than that of the light it is possible to "send out" an observer:

$$(\forall m \in \mathsf{Ob})(\forall p, q \in \mathsf{Q}^n)[\mathsf{space}(p,q) < \mathsf{time}(p,q) \rightarrow \\ (\exists k \in \mathsf{Ob})(\mathsf{W}(m,k,p) \wedge \mathsf{W}(m,k,q))].$$

In geometric terms this means that each line in the coordinate system with slant smaller than 1 is the worldline of a (potential) observer, in m's worldview.

The next axiom is the formalized version of postulate (iv) in Sec. 2.1.

Axiom 6 (AxEvent) If an observer observes three bodies at the same space-time location, then all other observers observe that these three bodies meet:

$$(\forall m, k \in \mathsf{Ob})(\forall b, b', b'' \in \mathsf{B})(\forall p \in \mathsf{Q}^n)(\exists p' \in \mathsf{Q}^n) \\ [\mathsf{W}(m,b,p) \wedge \mathsf{W}(m,b',p) \wedge \mathsf{W}(m,b'',p) \rightarrow \\ (\mathsf{W}(k,b,p') \wedge \mathsf{W}(k,b',p') \wedge \mathsf{W}(k,b'',p'))].$$

On AxEvent: In Sec. 2.1 we talked about "events". E.g. "Rear sent light signal Ph1" was called an event, another event was that "Nose sent light signal Ph2", and a third event was that "Middle, Ph1, and Ph2 meet". In the axiom AxEvent above, we talk about "3-meetings". We will reserve the word "event" for the set of all bodies present at a space-time location. Let us call $\mathsf{AxEvent}^+$ the axiom we obtain from AxEvent by replacing "3-meetings" with "events" in it, in this latter sense. We will see at the end of this subsection (cf. Theorem 11.2) that, in our approach, AxEvent is equivalent with this seemingly stronger axiom.

$\mathbf{Specrel}_0 := \{\mathsf{AxField}, \mathsf{AxSelf}, \mathsf{AxLine}, \mathsf{AxThEx}, \mathsf{AxEvent}, \mathsf{AxPh}\}$.

$\mathbf{Specrel}_0$ is the formalized version of SRK = NK^-+Light Axiom introduced in Sec. 2.1. Most of the interesting predictions of special relativity can be proved (in the rigorous manner of first-order logic) from $\mathbf{Specrel}_0$. However, some of the predictions have a little bit more complicated forms because different observers may use different "units of measurement". The last axiom brings the units of measurement of two observers to a common "platform".

For an observer m and space-time location $p \in \mathsf{Q}^n$, $\mathsf{ev}_m(p)$ denotes the "full event" happening in m's coordinate system at p,

$$\mathsf{ev}_m(p) := \{b \in \mathsf{B} : \mathsf{W}(m,b,p)\}.$$

We call the next axiom the *Axiom of Simultaneous Distance*.

Axiom 7(AxSim) Any two observers agree on the spatial distance between two events, if these two events are simultaneous for both of them:

$$(\forall m, k \in \mathsf{Ob})(\forall p, q, p', q' \in \mathsf{Q}^n)[\,\mathsf{ev}_m(p) = \mathsf{ev}_k(p') \wedge \mathsf{ev}_m(q) = \mathsf{ev}_k(q') \wedge \\ \mathsf{time}(p,q) = \mathsf{time}(p',q') = 0 \quad \rightarrow \quad \mathsf{space}(p,q) = \mathsf{space}(p',q')\,].$$

$\mathbf{Specrel} := \mathbf{Specrel}_0 \cup \{\mathsf{AxSim}\}$
$= \{\mathsf{AxField}, \mathsf{AxSelf}, \mathsf{AxLine}, \mathsf{AxThEx}, \mathsf{AxEvent}, \mathsf{AxPh}, \mathsf{AxSim}\}$.

In Sec. 2.5 we will prove that **Specrel** is consistent, and hence the weaker $\mathbf{Specrel}_0$ is also consistent. This will show that we have succeeded in eliminating the contradiction from (NK+Light Axiom): there is no statement A such that from the new theory (NK$^-$+Light Axiom) we can derive both A and its negation $\neg A$. In the next subsection we will prove that $\mathbf{Specrel}_0$ implies (in the rigorous manner of first-order logic) the negations of (v) and (iii), i.e. the negations of "absolute time" and "all motion is possible". In the next subsection we will also begin to investigate what the world looks like assuming SRK, in which ways it is different from our common-sense Newtonian world. Before doing this, we show two simple properties of $\mathbf{Specrel}_0$.

An important theme will be to establish which things all the observers perceive ("see") the same way, and which things they perceive differently. The things that they see the same way will be called "*absolute*", the things that they see differently will be called "*relative*". Whence the name "relativity theory". First we show that all observers see the same "events" to occur, and not only they see the same 3-meetings to occur.

Let $\mathsf{AxEvent}^+$ denote the statement that if an observer observes an event, then all other observers observe this event:

$$(\forall m, k \in \mathsf{Ob})(\forall p \in \mathsf{Q}^n)(\exists p' \in \mathsf{Q}^n)(\forall b \in \mathsf{B})[\mathsf{W}(m, b, p) \leftrightarrow \mathsf{W}(k, b, p')].$$

The symbol $\models$ denotes the *semantic consequence* relation of FOL. Before discussing the details, we note that in the case of FOL, $\models$ coincides with FOL-provability $\vdash$. If $\mathfrak{M}$ is a possible model and φ is a FOL formula, then $\mathfrak{M} \models \varphi$ abbreviates the statement "formula φ is valid in model $\mathfrak{M}$". For a set Ax of formulas, $Ax \models \varphi$ means that for every possible model $\mathfrak{M}$, if $\mathfrak{M} \models Ax$, then $\mathfrak{M} \models \varphi$.

THEOREM 11.2 $\{\mathsf{AxEvent}, \mathsf{AxPh}, \mathsf{AxField}\} \models \mathsf{AxEvent}^+$.

Proof Assume $\mathfrak{M} = \langle Q, \ldots, W\rangle \models \{\mathsf{AxEvent}, \mathsf{AxPh}, \mathsf{AxField}\}$ and let $m, k \in Ob$, $p \in Q^n$. There are (at least) two distinct lines ℓ_1, ℓ_2 of slope 1 going through p, e.g. $\ell_1 = \{p + x * \langle 1, -1, 0, \ldots, 0\rangle : x \in Q\}$ and $\ell_2 = \{p + x * \langle 1, 1, 0, \ldots, 0\rangle : x \in Q\}$ are such. (We used $\mathsf{AxField}$ here.) There are photons $\mathsf{ph}_1, \mathsf{ph}_2$ "living on ℓ_1, ℓ_2" respectively, by AxPh. (I.e. $\ell_i = \mathsf{wline}_m(\mathsf{ph}_i)$ for $i = 1, 2$.) Let us consider the worldlines of these photons in k's worldview. By AxPh, these are straight lines of slope 1. We are going to show that they intersect in a unique point.

We have that $\mathsf{wline}_m(\mathsf{ph}_1) \neq \mathsf{wline}_m(\mathsf{ph}_2)$. Let $q \in \ell_1$, $q \notin \ell_2$ and let ℓ_3 be the straight line of slope 1 going through q and parallel with ℓ_2. (I.e. $\ell_3 = \{q + x * \langle 1, 1, 0, \ldots, 0\rangle : x \in Q\}$.) Let $\mathsf{ph}_3 \in \mathsf{Ph}$ be such that $\mathsf{wline}_m(\mathsf{ph}_3) = \ell_3$.

Now, m "sees" 3-meetings of $\{\mathsf{ph}_1, \mathsf{ph}_2, \mathsf{ph}_2\}$, $\{\mathsf{ph}_1, \mathsf{ph}_3, \mathsf{ph}_3\}$ but m does not see a 3-meeting of $\{\mathsf{ph}_2, \mathsf{ph}_3, \mathsf{ph}_3\}$. By AxEvent, the same must hold for k. Thus $\mathsf{wline}_k(\mathsf{ph}_1)$ and $\mathsf{wline}_k(\mathsf{ph}_2)$ must meet but must not coincide and hence they intersect in a unique point.

Let p' be their intersection point, i.e. $\{p'\} = \mathsf{wline}_k(\mathsf{ph}_1) \cap \mathsf{wline}_k(\mathsf{ph}_2)$. Now, it is easy to show by using AxEvent again that $\mathsf{ev}_m(p) = \mathsf{ev}_k(p')$. (Indeed, let $b \in B$ be arbitrary. Then m sees a 3-meeting of $\mathsf{ph}_1, \mathsf{ph}_2, b$ iff $b \in \mathsf{ev}_m(p)$, and the same for m, p replaced with k, p'.) QED

THEOREM 11.3 *No observer observes the same event at two different space-time locations in models of* $\{\mathsf{AxPh}, \mathsf{AxField}\}$.

Proof Let $p, q \in \mathsf{Q}^n$, $p \neq q$. There is a straight line ℓ of slope 1 through p which avoids q (because through each point p there are at least 2 distinct lines of slope 1). By AxPh, ℓ is the worldline of a photon $\mathsf{ph} \in \mathsf{Ph}$ (in m's worldview). Then $\mathsf{ph} \in \mathsf{ev}_m(p)$ while $\mathsf{ph} \notin \mathsf{ev}_m(q)$, showing that $\mathsf{ev}_m(p) \neq \mathsf{ev}_m(q)$. QED

Theorems 11.2 and 11.3 imply that in each observer's worldview, the space-time locations and the events observed by any observer are in one-one correspondence. Thus, in a given observer's worldview, we can speak of events as if they were space-time locations. E.g. we can quantify over events, meaning that we have in fact quantified over space-time locations. By the same token, we can apply any function defined on space-time locations to events. Specifically, let $\mathsf{loc}_m(e)$ denote the *location of event* e in m's worldview, then

$$\mathsf{loc}_m(e) = p \quad \text{iff} \quad \mathsf{ev}_m(p) = e.$$

By the *time-distance between two events as seen by an observer* we will mean the time-distance between the space-time locations where the observer sees the two events, and similarly for *spatial distance*. Formally, with $p = \mathsf{loc}_m(e)$, $p' = \mathsf{loc}_m(e')$ we have

$$\mathsf{time}_m(e, e') := \mathsf{time}(p, p'), \qquad \mathsf{space}_m(e, e') := \mathsf{space}(p, p'),$$

$\mathsf{time}_m(e) := p_1$ denotes the time where m sees event e happen, and $\mathsf{space}_m(e) := \langle 0, p_2, \ldots, p_n \rangle$ denotes the space-location where m sees event e happen.

2.4 Characteristic differences between Newtonian and special relativistic kinematics

The most frequently quoted predictions of special relativity are the following three *paradigmatic effects*. (1) moving clocks slow down, (2) moving meter-rods shrink, and (3) moving pairs of clocks get out of synchronism. These three effects are easily formulated in the first-order language introduced so far.

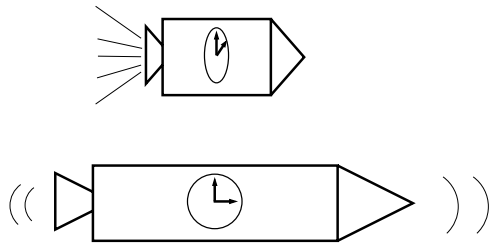


Figure 11.5. Moving clocks slow down and moving spaceships shrink.

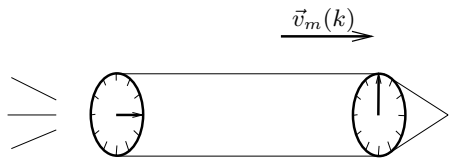


Figure 11.6. Moving clocks get out of synchronism.

Let m, k be observers in a model of our language. By the *direction of spatial separation* of two events e, e' in m's worldview we mean the natural thing, i.e. we mean the straight line connecting the "spatial projections" $\langle 0, p_2, \ldots, p_n\rangle$ and $\langle 0, q_2, \ldots, q_n\rangle$ if p and q are the space-time locations m sees e and e' at, respectively (or the point $\langle 0, p_2, \ldots, p_n\rangle$ if these two points are the same). The *spatial direction of motion* of a body b in m's worldview is the direction of spatial separation of two distinct events in $\mathsf{wline}_m(b)$, whenever the latter is a straight line. (In order to deal with the "degenerate" situations in the next theorem, we say that a point is both parallel and orthogonal to a line or to another point.) We say that e, e' are *simultaneous* in m's worldview iff $\mathsf{time}_m(e, e') = 0$. Let $v_m(k)$ denote the *speed of* k as seen by m, i.e. $v_m(k)$ is the slope of the worldline of k in m's worldview. We note that $\mathbf{Specrel}_0 \not\models (\forall m, k \in \mathsf{Ob})v_m(k) = v_k(m)$ while $\mathbf{Specrel} \models (\forall m, k \in \mathsf{Ob})v_m(k) = v_k(m)$ (see Corollary 11.13).

Theorem 11.4 below implies that Absolute Time (i.e. (v) of NK) is inconsistent with SRK (i.e. with NK$^-$+Light Axiom), hence it was necessary to omit it from NK. Theorem 11.4 says that simultaneity of events is not absolute. Actually, it implies something more surprising, more exotic: the question of what happened earlier and what later is not absolute either (see Corollary 11.5 after the theorem). Fig. 11.7 illustrates the statements in Theorem 11.4. Theorem 11.4 is a more detailed version of Proposition 11.1.

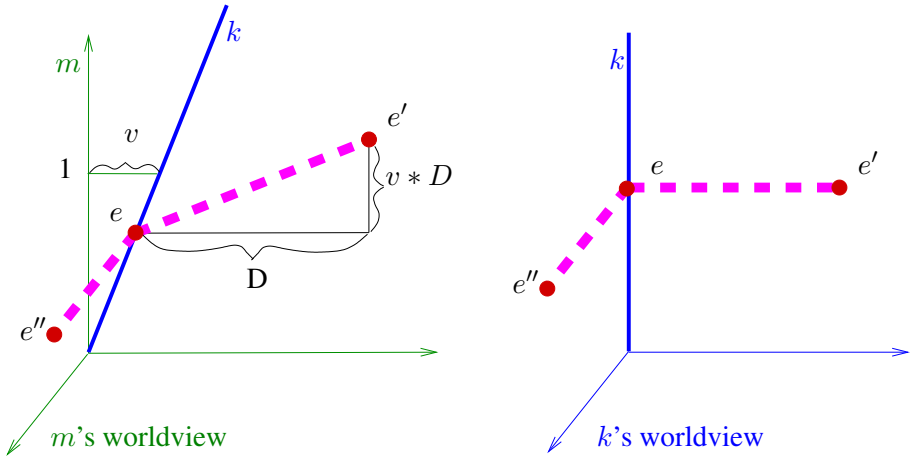


Figure 11.7. Illustration for Theorem 11.4. Simultaneity of events is not absolute. Events e, e', e'' are simultaneous for k, but e, e' are not simultaneous for m.

THEOREM 11.4 (SIMULTANEITY OF EVENTS IS NOT ABSOLUTE) *Assume* **Specrel**$_0$ *and let* m, k *be observers. Statements (i) and (ii) below hold.*

(i) *Assume that in* m*'s worldview the spatial separation of events* e, e' *is parallel with the direction of motion of* k*. Then*

e, e' *are simultaneous in* k*'s worldview*

iff

$\mathsf{time}_m(e, e') = v_m(k) * \mathsf{space}_m(e, e')$.

(ii) *Assume that* e, e'' *are simultaneous both in* k*'s worldview and in* m*'s worldview. Then in* m*'s worldview the spatial separation of* e, e'' *is orthogonal to the direction of motion of* k*.*

Proof The proof of Theorem 11.4 follows the structure and ideas of the intuitive proof of Proposition 11.1.

Proof of (i). Let e, e' be simultaneous events in k's worldview. Let $p = \mathsf{loc}_k(e)$, $p' = \mathsf{loc}_k(e')$ and let $q = (1/2) * (p + p')$, their "middle-point". Let ℓ_1, ℓ_2, ℓ_3 be straight lines parallel with the time-axis and going through p, p', q respectively and let m_1, m_2, m_3 be observers with $\ell_i = \mathsf{wline}_k(m_i)$ $(i = 1, 2, 3)$. See Fig. 11.8.

Let $\delta := |p - q| = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$. This exists since the field $\langle Q, +, *, \leq \rangle$ is quadratic by $\mathsf{AxField}$. By using δ now we can construct straight lines of slope 1 connecting ℓ_1, ℓ_2, ℓ_3 as follows. Let $u = \langle 1, 0, \ldots, 0 \rangle$

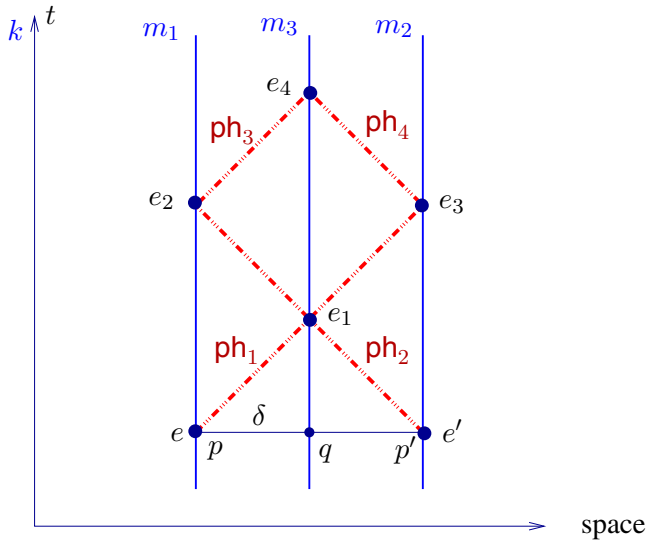


Figure 11.8. Illustration for the proof of Theorem 11.4(i).

(the "time unit-vector"), $p1 = q + \delta * u$, $p2 = p + 2\delta * u$, $p3 = p' + 2\delta * u$, $p4 = p1 + 2\delta * u$. The straight lines connecting the points $pp1$, $p'p1$, $p2p4$, and $p3p4$ all have slope 1 by construction, hence by AxPh there are photons $\mathsf{ph}_1, \ldots, \mathsf{ph}_4$ whose worldlines these are, respectively. Let $e_i = \mathsf{ev}_k(pi)$ for $i = 1, \ldots, 4$. See Fig. 11.8.

We will think of the pattern constructed so far as representing the two thought-experiments in Proposition 11.1. We will think of k, m_1, m_2, m_3 as the Spaceship, Rear, Nose, and Middle respectively; e is the event when Rear sent ph_1 towards Middle, e' is the event when Nose sent ph_2 towards Middle, and e_1 is the event when these two reached Middle. The upper part of the arrangement (events $e_1, \ldots, e_4$ and photons $\mathsf{ph}_1, \ldots, \mathsf{ph}_4$) represents the experiment of Middle by which he tested that he indeed was standing in the middle (the two photons $\mathsf{ph}_2, \mathsf{ph}_1$ sent towards Rear and Nose arrived back, after bouncing back at the mirrors, at the same time in event e_4).

Switch now to the worldview of m! See Fig. 11.9. The worldlines of $m_1, m_2, m_3, \mathsf{ph}_1, \ldots, \mathsf{ph}_4$, respectively, are all straight lines, the last four of slope 1, by AxLine, AxPh. The meeting points of these lines are exactly as those of the corresponding worldlines in k's worldview, by AxEvent. The worldlines of m_1, m_2, m_3, k are parallel in m's worldview because they are so in k's worldview. (In more detail, e.g. for m_1, m_2: their worldlines do not meet, by AxEvent. Their worldlines are not skew, because one can construct photons

$\mathsf{ph}_1, \mathsf{ph}_2$ with meeting points e, e', e_1, e_2, e_3 as in Fig. 11.8, and this ensures that they are in one plane (i.e. in the plane determined by e, e', e_1).) Now assume that the spatial separation of e, e' is parallel with the spatial direction of movement of k, in m's worldview. This means that there is a plane P containing the whole configuration (the worldlines of $m_1, \ldots, \mathsf{ph}_4$), and it is *vertical*, i.e. it contains a line parallel with the time-axis. The worldlines of $\mathsf{ph}_1, \mathsf{ph}_2$ are parallel with

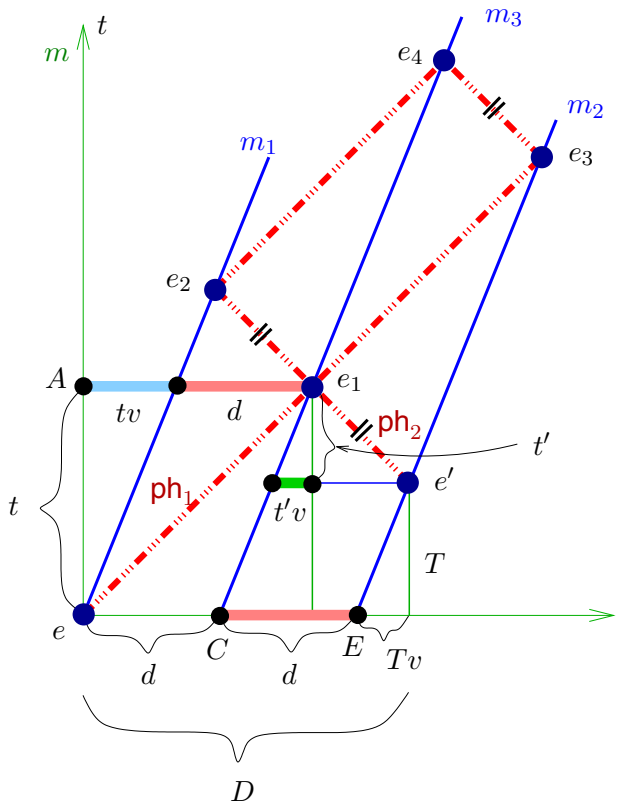


Figure 11.9. Illustration for the proof of Theorem 11.4(i).

those of $\mathsf{ph}_3, \mathsf{ph}_4$, respectively, in m's worldview, because they are so in k's worldview (and because they are all in one plane). Thus the distance of events e_2 and e_1 is the same as the distance of events e_4 and e_3 according to m, i.e. with the notation $r_i = \mathsf{loc}_m(e_i)$ we have $|r_2 - r_1| = |r_4 - r_3|$. Similarly, since the lines connecting e_4, e_1 and e_3, e' are also parallel, we get $|r_1 - r'| = |r_3 - r_4|$ (where $r' = \mathsf{loc}_m(e')$). Thus $|r_2 - r_1| = |r_1 - r'|$. For this reason, the distance between e and C in Fig. 11.9 is the same as the distance between C and E,

i.e. C is the middle-point of e and E. Thus m also sees that m_3 is positioned exactly in the middle of m_1 and m_2.

We now recall the intuitive chain of thought in the proof of Proposition 11.1 taking into account the quantitative aspects, too. We think of m as Earth. Let us measure time in "seconds", and let us assume that the length of the ship as Earth sees it is $2d$. Now, Earth sees that inside the spaceship the photon ph_1 covered distance d with velocity $1 - v$, thus it took $t = d/(1 - v)$ seconds for ph_1 to reach the middle of the ship. Similarly ph_2 covered distance d with velocity $1 + v$, so it took $t' = d/(1 + v)$ seconds for it to reach the middle of the ship. Hence Nose had to send ph_2 exactly $T = t - t' = [d/(1 - v)] - [d/(1 + v)] = 2dv/(1 - v^2)$ seconds later in order that they meet in the middle. Since the photon was sent T seconds later, the ship covered $T * v$ distance during this time, thus the spatial distance of event e (which is sending out photon ph_1) and event e' (which is sending out photon ph_2) is $D = 2 * d + T * v = ([2d(1 - v^2)] - [2dv^2])/(1 - v^2) = 2d/(1 - v^2)$. Hence $T = v * D$, as was to be shown. This computation can be faithfully reconstructed in the settings of the present Theorem 11.4, see Fig. 11.9.

This proves the "only if" part in (i). The "if" part in (i) can be proved by taking an event e'' in k's worldview which is simultaneous with e and which takes place at the same place as e', i.e. $\mathsf{space}_k(e'', e') = 0$; now we can use the previously proven part for e, e'' and then use $\mathsf{time}_m(e'', e') \neq 0$.

The **proof of (ii)** is similar to that of (i), we include Fig. 11.10 for illustration.

QED

Remark. The first-order logic axiomatization of relativity theories we are describing here is a very good place for applying Tarski's first-order logic axiomatization of Euclidean geometry. We try to illustrate this claim. In the proof of Theorem 11.4 we used several geometrical properties of the Euclidean geometry G built on the field $\langle Q, +, * \rangle$ in place of the reals. E.g. we used that "for any two distinct points there is a unique (straight) line connecting them", or "through any point there is a unique line ℓ' parallel with ℓ", we used the notions of planes, being parallel etc. The properties of G we used in the proof are easy to check directly by using the axioms of a quadratic ordered field and the (analytic) definitions of a straight line etc. There is another way, though. Instead of directly checking in the geometry G validity of each statement which arises in the proof, we could use the axioms in a first-order logic axiomatization of (synthetic) Euclidean geometry and derive everything from those axioms (or just rely on the existing theorems and definitions of this area of research). We recall that Hilbert, 1977 axiomatized Euclidean geometry over the reals by using second-order logic axioms, and Tarski, 1959 gave a first-order logic axiom system for this geometry. This also made possible to replace the field of reals with arbitrary fields and investigate what algebraic properties of the field correspond to what geometrical properties. This subject—which is highly relevant in the

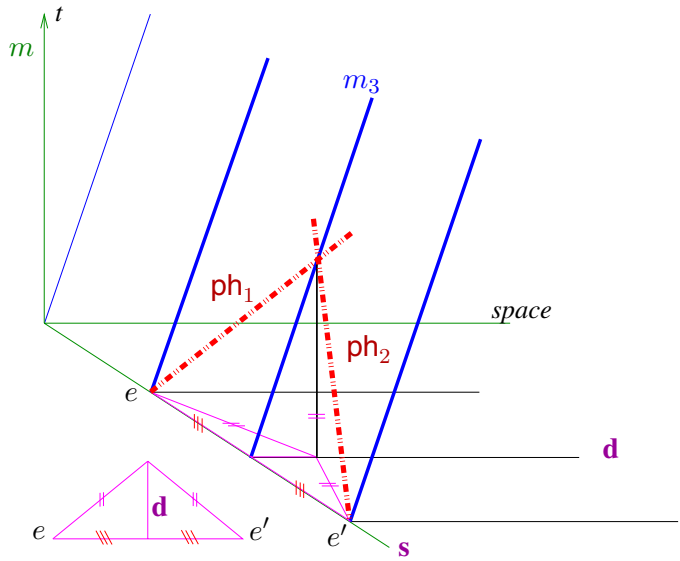


Figure 11.10. Illustration for the proof of Theorem 11.4(ii). If $\mathsf{ph}_1, \mathsf{ph}_2, m_3$ meet, then the spatial separation s of e, e' must be orthogonal to the spatial direction d of movement. This figure shows how m "sees" the thought-experiment illustrated in Fig. 11.8, but conducted in a spatial direction not necessarily parallel with the direction d of motion of k.

approach presented in this paper—is quite rich, see e.g. Tarski, 1959, Goldblatt, 1987, Schwabhäuser et al., 1983, Szczerba, 1970, Szmielew, 1974 and Aiello and van Benthem, 2002. Ax, 1978, Goldblatt, 1987 and Mundy, 1986 make use of Tarski's axiom system for Euclidean geometry in their axiomatizations of Special Relativity Theory.

Actually, Fig. 11.11 shows that by using the methods of axiomatic Euclidean geometry, the proof of Theorem 11.4 could be made simpler and more transparent. QED

COROLLARY 11.5 (THE TEMPORAL ORDER OF EVENTS IS NOT ABSOLUTE) *Assume* **Specrel**$_0$. *For all observers m, k not at rest relative to each other there are events e, e' such that e happens earlier than e' according to m while e happens later than e' according to k. Formally:*

$$\mathbf{Specrel}_0 \models (\forall m, k \in \mathsf{Ob})[\, v_m(k) \neq 0 \rightarrow (\exists e, e')[\, \mathsf{time}_m(e) < \mathsf{time}_m(e') \wedge \mathsf{time}_k(e) > \mathsf{time}_k(e')\,]\,].$$

Proof Assume that $v_m(k) \neq 0$. Let e, e' be distinct events in the life of k (i.e. $k \in e \cap e'$, $e \neq e'$) and assume $\mathsf{time}_m(e) < \mathsf{time}_m(e')$. If $\mathsf{time}_k(e) > \mathsf{time}_k(e')$ then we are done. So assume $\mathsf{time}_k(e) < \mathsf{time}_k(e')$. Let P be

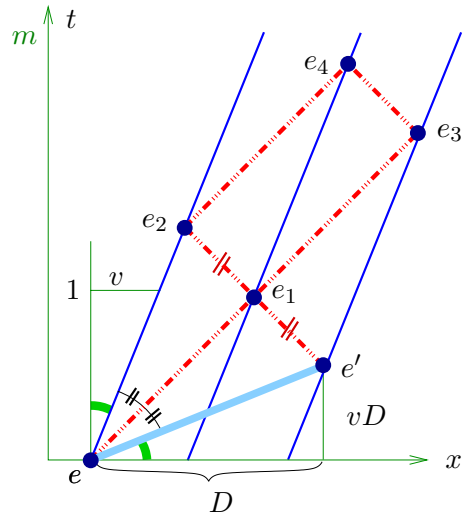


Figure 11.11. Simpler proof for Theorem 11.4(i) by using synthetic geometry: By using the parallelograms $e_2 - e_1 - e_3 - e_4$ and $e_1 - e' - e_3 - e_4$ we get that the distance between e_2 and e_1 is the same as the distance between e_1 and e'. The triangles $e - e_2 - e_1$ and $e - e' - e_1$ are congruent since the two photon-lines are orthogonal to each other. Since the slope of the light-line $e - e_1$ is 1, the angle between the time-axis $\overline{t}$ and $e - e_2$ is therefore the same as the angle between the "space axis" $\overline{x}$ and $e - e'$, yielding the desired result.

the "plane of movement of k", i.e. let P be a plane parallel to the time-axis and which contains $\mathsf{wline}_m(k)$. By Theorem 11.4(i), the events on P which are simultaneous according to k with e form a straight line ℓ which is not "horizontal", see Fig. 11.12. Therefore there is an event e'' on ℓ such that $\mathsf{time}_m(e'') > \mathsf{time}_m(e')$. Now $\mathsf{time}_k(e'') = \mathsf{time}_k(e) < \mathsf{time}_k(e')$, and we are done. QED

In the next theorem we formalize the three paradigmatic effects of SRK and prove them from **Specrel**. Figs. 11.13, 11.14 illustrate the statement of Theorem 11.6 in cartoon and in space-time diagram respectively, while Fig. 11.17 at the end of the proof summarizes in one picture how two observers "see" each other's coordinate systems. Fig. 11.18 gives a geometric illustration and explanation for the three paradigmatic effects.

THEOREM 11.6 (THE THREE PARADIGMATIC EFFECTS) *Assume* **Specrel** *and $n \geq 3$, and let m, k, k' be observers with $v := v_m(k)$, and $v_k(k') = 0$. Assume that k's spaceship, the rear and nose of which are marked by observers k, k', moves forwards (i.e. $\mathsf{wline}_m(k), \mathsf{wline}_m(k')$ are contained in a plane parallel with the time-axis and for some e'', e^* with $k \in e''$, $k' \in e^*$*

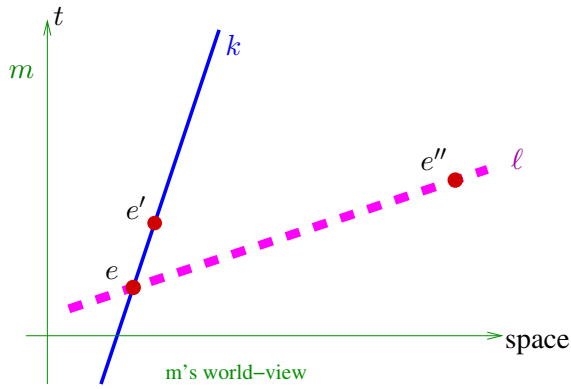


Figure 11.12. Illustration for the proof of Corollary 11.5. In m's worldview, e happens earlier than e' and e'' happens later than e'. However, in k's worldview e is simultaneous with e''.

and $\mathsf{space}_m(e'', e^*) = 0$ *we have* $\mathsf{time}_m(e^*) < \mathsf{time}_m(e'')$*), time flows forwards for* k *as seen by* m *(i.e.* $\mathsf{time}_m(\mathsf{ev}_k(\overline{0})) < \mathsf{time}_m(\mathsf{ev}_k(\mathbf{1}_t))$*). Assume further that, according to* k*, the clocks in the ship are synchronized (i.e.* $\mathsf{time}_k(\mathsf{ev}_{k'}(\overline{0})) = 0$*) and the length of the ship is* D *(i.e.* $|\mathsf{space}_k(e')| = D$*), see Fig. 11.14). Then (1)–(3) below hold.*

(1) *(moving pairs of clocks get out of synchronism) According to* m*, the clock-readings at the nose of* k*'s spaceship are* $v * D$ *less than the simultaneous readings at the rear of the ship. (The clocks in the nose are late relative to those in the rear. See Fig. 11.13.) Formally:*

$$(\forall e, e')[\, k \in e \ \wedge\ k' \in e' \ \wedge\ \mathsf{time}_m(e, e') = 0 \rightarrow \\ \mathsf{time}_{k'}(e') = \mathsf{time}_k(e) - (v * D)\ ].$$

(2) *(moving clocks slow down (called "time-dilation")) Any process that lasts* t *seconds in the ship, lasts for* $t/\sqrt{1-v^2}$ *seconds as seen by* m*. Formally:*

$$(\forall e, e')[\, k \in e \cap e' \rightarrow \mathsf{time}_m(e, e') = \mathsf{time}_k(e, e')/\sqrt{1-v^2}\,].$$

(3) *(moving ships get shorter (called "length-contraction")) According to* m*, the length of* k*'s ship is only* $D * \sqrt{1-v^2}$ *(and not* D *as* k *states). Formally:*

$$(\forall e, e')[\, k \in e \wedge k' \in e' \wedge \mathsf{time}_m(e, e') = 0 \rightarrow \\ \mathsf{space}_m(e, e') = D * \sqrt{1-v^2}\,].$$

Hence, moving spaceships become 'squat": They get shorter but their width and height do not change by AxSim*. So they get distorted. This distortion effect remains true in weaker fragments of* **Specrel***, e.g. in*

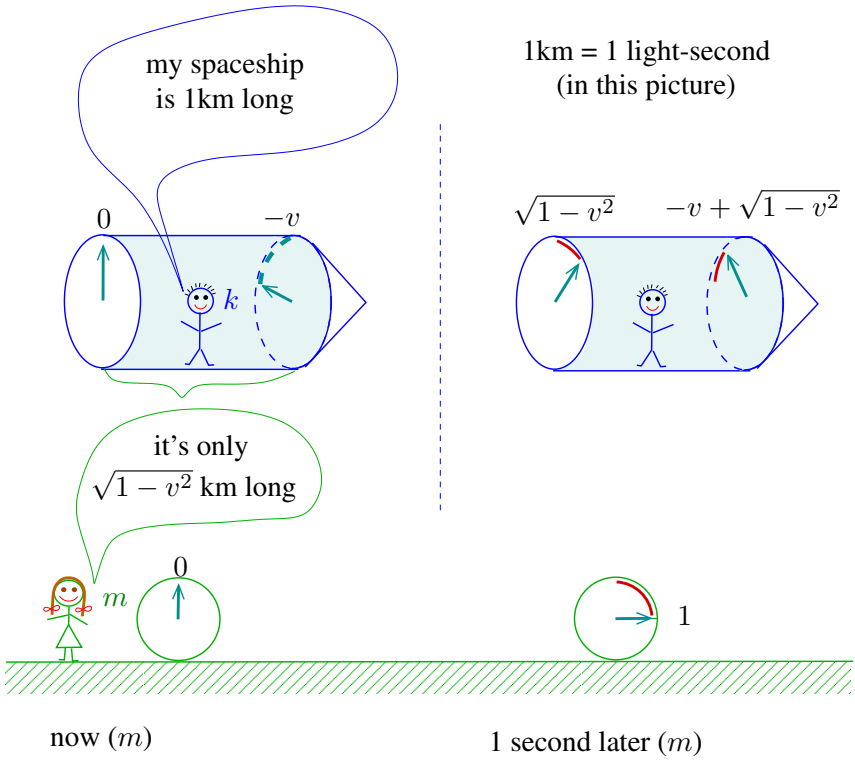


Figure 11.13. Illustration for Theorem 11.6. According to m, the length of the spaceship is d km, it is 1 km wide and tall, and the clocks in the nose show $dv/\sqrt{1-v^2}$ less time than those in the rear. According to k, the length of the ship is $D = d/\sqrt{1-v^2}$, it is 1 km wide and tall, and the clocks in the nose and the ones in the rear all show the same time. In the picture we chose $d = \sqrt{1-v^2}$ and $D = 1$. Compare this picture with Fig. 11.3. As v increases, the spaceship becomes squat: it becomes shorter while its width and height remain the same. Cf. also Figs. 11.5, 11.6.

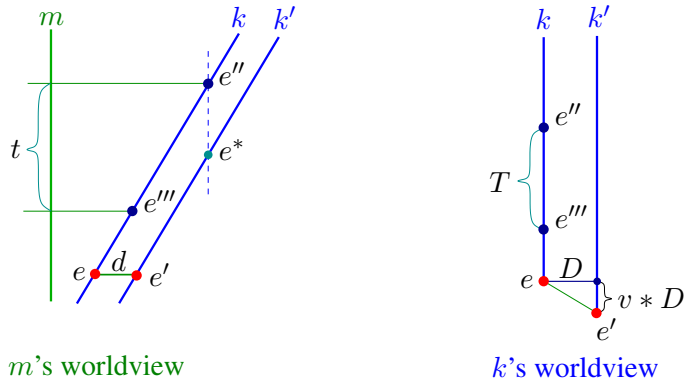


Figure 11.14. Illustration for Theorem 11.6, which states that $t = T/\sqrt{1-v^2}$ and $d = D * \sqrt{1-v^2}$.

Specrel$_0$, *and for arbitrary* n*, as shown in Andréka et al., 2002, Sec. 4.8, esp. p. 653.*

Proof To prove Theorem 11.6, we will use two new thought-experiments analogous to the one in Proposition 11.1. Indication for the proof in geometrical flavor is in the caption of Fig. 11.18. Let m, k, k', v be as in the hypothesis part of the theorem.

The thought-experiment for proving time-dilation uses *Einstein's light-clock*, cf. Fig. 11.15. This light-clock consists of two mirrors and a photon which bounces back and forth between the two mirrors. The two mirrors, M_1 and M_2, are positioned at the rear of k's spaceship so that their spatial separation is orthogonal to the movement of the ship (as seen by m) and their spatial distance is 1 light-second. Thus for the photon ph from one mirror M_1 to the other M_2 lasts for 1 second; one tick of the clock lasts 1 second as seen from the ship k. Let e, e' be the events when ph leaves mirror M_1 and reaches mirror M_2, respectively. According to m, the spatial distance between e and e' is not 1 (as seen from k's ship) but $\sqrt{1+x^2}$ where x is the distance the second mirror M_2 covers while the photon reaches it. If t is the time elapsed between e and e' as m sees it, then $x = t * v$, and thus $\mathsf{space}_m(e, e') = \sqrt{1+x^2}$, hence $t = \sqrt{1+x^2}$ because the speed of ph is 1 in m's worldview, too. Now from $t^2 = 1 + t^2v^2$ we get $t = 1/\sqrt{1-v^2}$ (which is greater than 1). We obtained the desired rate of time-dilation.

The thought-experiment for proving length-contraction (Theorem 2.4(3)) uses a so-called "two-dimensional light-clock", this is the following. See Fig. 11.16. There are two pairs of mirrors and two photons bouncing between them. The first pair of mirrors M_1, M_2 and the photon ph bouncing between

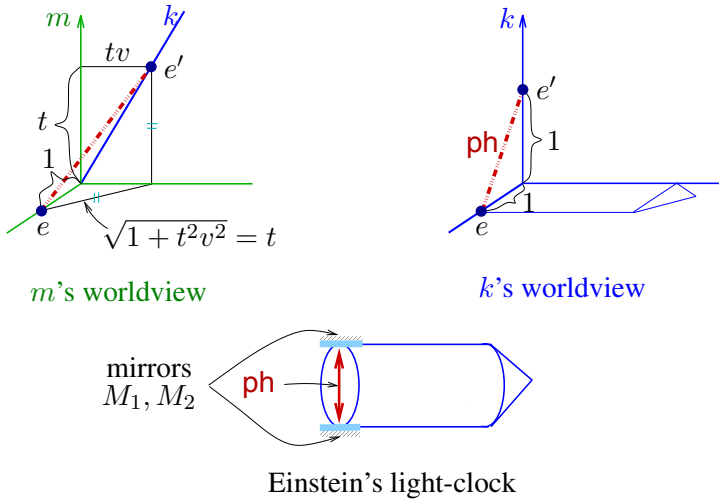


Figure 11.15. Illustration for the proof of Theorem 11.6(2) (time-dilation). Einstein's light-clock consists of two mirrors and a photon ph bouncing between them. One tick lasts $t = (1/\sqrt{1-v^2})$ seconds in m's worldview, if one tick lasts 1 second in the ship.

them is as in Einstein's light-clock. The second pair of mirrors M_3, M_4 and ph' are like M_1, M_2, ph with the difference that M_3, M_4 are separated in the direction of movement of the ship. Thus if $\mathsf{ph}, \mathsf{ph}'$ leave mirrors M_1, M_3 in the same event e (we may assume that M_3 is positioned where M_1 is), then after bouncing they will be back in the same event e' again. The whole scene in m's worldview is as follows. Photon ph behaves exactly as in Einstein's light-clock, so $t = \mathsf{time}_m(e, e') = 2/\sqrt{1-v^2}$, as before. Let us see what the "tick" made by ph' looks like in m's worldview. By the arguments in the proof of Theorem 11.4(i), if $d = \mathsf{space}_m(e, e')$, then $t = d/(1-v) + d/(1+v) = 2d/(1-v^2)$. Hence $d = \sqrt{1-v^2}$, the distance between mirrors M_3, M_4 is $\sqrt{1-v^2}$ (which is smaller than 1) as seen by m and not 1 as seen by k. We obtained the desired rate of length-contraction.

We get Theorem 11.6(1) by combining Theorem 11.4(i) and Theorem 11.6(2) (cf. e.g. Fig. 11.17). QED

Theorem 11.7 below implies that the statement "Motion with every finite speed is possible" (i.e. (iii) of NK) is inconsistent with $\mathbf{Specrel}_0$. This justifies the step of weakening (iii) to (iii)$^-$ in NK$^-$. Theorem 11.7 below also shows that we do not have to postulate as an axiom that no observer can move faster than light; as an axiom this would be difficult to motivate. Luckily, "no faster-than-light observer" turns out to be a corollary of the well-motivated axioms in

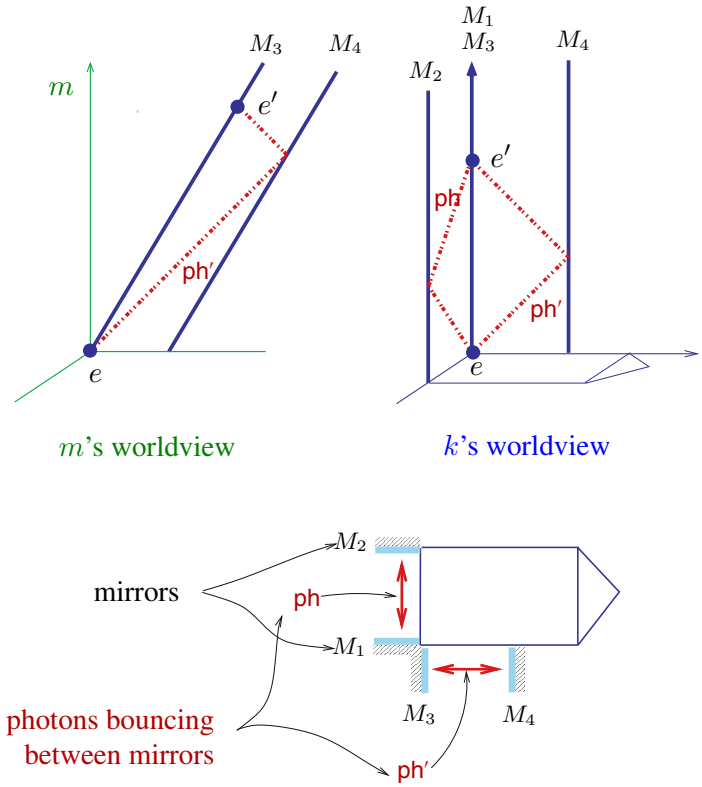


Figure 11.16. Illustration for the proof of Theorem 11.6(3) (length-contraction). The "two-dimensional" light-clock consists of two pairs of mirrors and two photons (ph, ph′) bouncing between them. The two photons' bouncing-time is the same in k's worldview, thus it has to be the same in m's worldview, too.

Specrel$_0$. Putting it more succinctly: "no faster-than-light observer" (No FTL for short) is a theorem only in our approach and not an axiom.

THEOREM 11.7 (NO FASTER-THAN-LIGHT MOTION) *Assume* **Specrel**$_0$.

(i) *No observer can move with the speed of light, i.e.* $v_m(k) \neq 1$ *for all observers* m, k.

(ii) *Assume* $n \geq 3$. *Then* $v_m(k) < 1$ *for all observers* m, k, *i.e. if* $n > 2$ *then no observer can move faster than light. If* $n = 2$, *then* $v_m(k) > 1$ *for some observers* m, k *is possible.*

For proof see e.g. Andréka et al., 1999, Proposition 1, Theorem 3, Madarász, 2002, 2.3.5, 2.8.25, 3.2.13, Madarász et al., 2004, Theorem 3, Theorem 5. A proof can also be reconstructed from the proof of Theorem 11.11, p. 641. QED

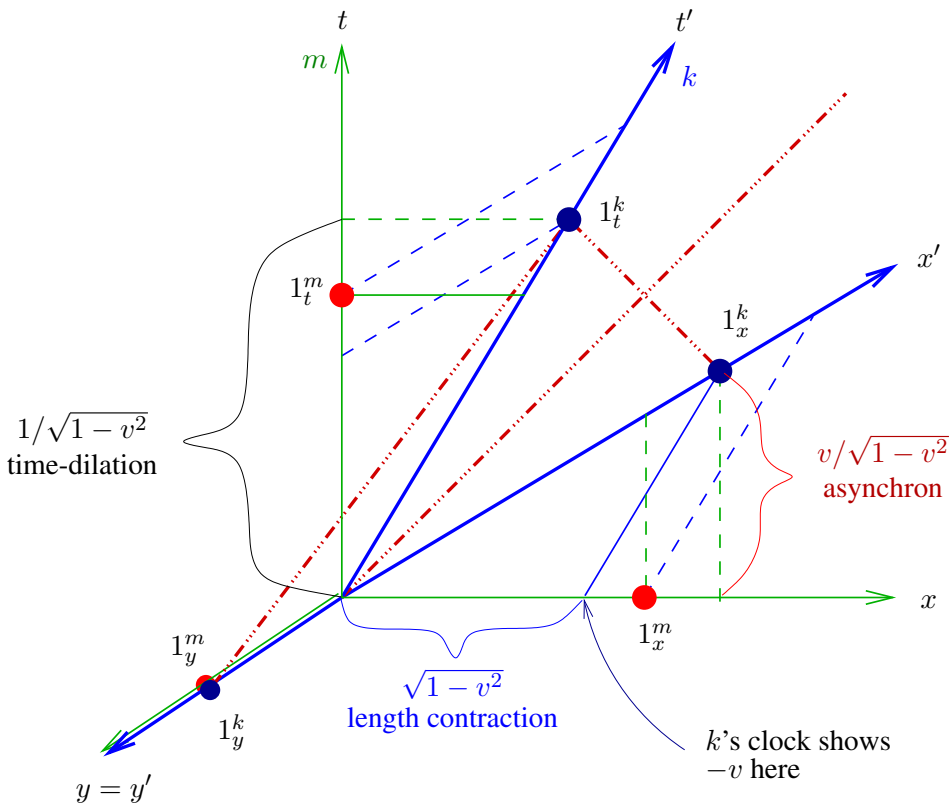


Figure 11.17. Illustration for Theorem 11.6. The three photon-worldlines illustrate, in some sense, the three thought-experiments for proving time-dilation, length-contraction, and getting out of synchronism, respectively. This figure also illustrates Lorentz transformations (Def. 11.9).

For a conceptual analysis of the above No FTL theorem we refer to Sec. 2.7. There we will address the question "why No FTL?", e.g. which parts of $\mathbf{Specrel}_0$ are responsible for No FTL, etc.

In Newtonian Kinematics, NK, spatial distance of events is not absolute (e.g. two events that took place in the dining car of a moving train at different times, took place at different places for someone not on the train), but the time elapsed between two events is the same for any two observers, moving relative to each other or not. Theorem 11.6(ii) says that in SRK the time elapsed between two events is not absolute, either. In this respect, SRK is a more symmetric theory than NK. But is there anything left that the observers see the same way? Curiously, a "mix" of time and space does remain absolute (as opposed to being relative like time and space are).

THEOREM 11.8 (RELATIVISTIC DISTANCE) *Assume* **Specrel** *and* $n \geq 3$, *and let* m, k *be observers,* e, e' *be events. Then*

$$\mathsf{time}_m(e,e')^2 - \mathsf{space}_m(e,e')^2 = \mathsf{time}_k(e,e')^2 - \mathsf{space}_k(e,e')^2. \qquad \textit{QED}$$

The above theorem is the starting point for building Minkowski geometry, which is the "geometrization" of SRK. It also indicates that time and space are intertwined in SRK.

Let us denote the quantity that is the same for all observers as stated in Theorem 11.8 above by

$$\boldsymbol{\mu}(p,q) := \mathsf{time}(p,q)^2 - \mathsf{space}(p,q)^2 \ .$$

The letter $\boldsymbol{\mu}$ refers to *Minkowski distance* (also called *relativistic distance*). We will see that every coordinate property that the observers observe the same way about events (in a model of **Specrel**) can be defined from this relativistic distance $\boldsymbol{\mu}$ (Corollary 11.17). More importantly, the whole structure $\mathfrak{M}$ can be retrieved from relativistic distance $\boldsymbol{\mu}$ (provided we disregard irrelevant properties of observers and photons like e.g. "there are several distinct photons on ph's worldline"). This means that we can re-define photons, observers, and even the field-operations on quantities Q from knowing only relativistic distance $\boldsymbol{\mu}$ (Theorem 11.16). This indicates that there is an "observer-independent reality" which is behind the different worldviews of the observers. This observer-independent reality is often called "objective" or absolute (as opposed to being "subjective" or relative like relative motion). We will explore these ideas in Sec. 2.6.

2.5 Explicit description of all models of Specrel, basic logical investigations

An advantage of having axiomatic theories is that we can use the benefits of the well-developed syntax-semantics duality of first-order logic. Namely, if we want to check whether a formula φ follows from **Specrel**, instead of making a rigorous syntactic derivation, we can check whether in all models of **Specrel** the formula φ holds or not. Specifically, we can *prove* that φ does not follow from **Specrel** by exhibiting a model of **Specrel** in which φ fails. In this subsection we give an explicit description of all models of **Specrel** and $\mathbf{Specrel}_0$. Based on this, then we will give a sample of logical investigations such as consistency, completeness, categoricity, decidability, and independence of axioms.

Theorems 11.4–11.7 in the previous subsection provide all the important ingredients for describing the models of our theories $\mathbf{Specrel}_0$ and **Specrel**. The "heart" of this description is the description of the so-called worldview transformations. Let m, k be observers. The *worldview transformation* w_{mk} relates the worldview of m with that of k, it relates those space-time locations where m and k observe the same events. I.e.

$$\mathsf{w}_{mk} := \{\langle p, q\rangle \in \mathsf{Q}^n \times \mathsf{Q}^n : \mathsf{ev}_m(p) = \mathsf{ev}_k(q)\}.$$

The worldview transformation is defined to be a binary relation on space-time locations, but under very mild assumptions it is a transformation of Q^n indeed and $\mathsf{ev}_k = \mathsf{ev}_m \circ \mathsf{w}_{km}$, hence the name "worldview transformation". (Here, and later, $f \circ g$ denotes the *composition* of functions f and g, i.e. $(f \circ g)(x) = f(g(x))$.) In fact, the worldview transformation $\mathsf{w}_{mk} : \mathsf{Q}^n \to \mathsf{Q}^n$ is the natural coordinate-transformation between the coordinate systems of m and k. It shows how the worldview of one observer m is distorted in the eye of another observer k.

Theorems 11.4, 11.6 give quite a lot of information on the worldview transformations in models of **Specrel**. They imply that w_{mk} is a Lorentz transformation as defined below, up to a suitable choice of coordinate directions.

It will be convenient to use the so-called unit-vectors. Let $1 \le i \le n$. The *i-th unit-vector* is

$$\mathbf{1}_i := \langle 0, \ldots, 0, 1, 0 \ldots, 0\rangle \quad \text{where the } 1 \text{ stands in the } i\text{-th place.}$$

We will also use the names $\mathbf{1}_t, \mathbf{1}_x, \mathbf{1}_y, \mathbf{1}_z$ for the first four unit-vectors. *From now on we fix a quadratic ordered field* $\mathfrak{Q} = \langle Q, +, *, \le\rangle$.

DEFINITION 11.9 (LORENTZ TRANSFORMATION) *Let* $-1 < v < 1$, $v \in Q$. *By the* Lorentz *transformation (or boost) of velocity* v *and over* $\mathfrak{Q}$ *we understand a linear mapping* $f : \mathsf{Q}^n \to \mathsf{Q}^n$ *for which*

$$f(\mathbf{1}_t) = \langle 1/\sqrt{1-v^2},\ v/\sqrt{1-v^2},\ 0, \ldots, 0\rangle,$$

$$f(\mathbf{1}_x) = \langle v/\sqrt{1-v^2},\ 1/\sqrt{1-v^2},\ 0,\ldots,0\rangle, \qquad \textit{and}$$

$$f(\mathbf{1}_i) = \mathbf{1}_i \quad \textit{for all } 3 \leq i \leq n.$$

Fig. 11.17 illustrates Lorentz transformations. A Lorentz transformation as a coordinate transformation usually is written as

$$t' = (t - vx)/\sqrt{1-v^2}, \quad x' = (x - vt)/\sqrt{1-v^2}, \quad y' = y, \quad z' = z.$$

The usual Newtonian (or Galilean) coordinate transformation is $t' = t$, $x' = x - vt$, $y' = y$, $z' = z$. Comparing the two transformations reveals that in SRK time and space are treated in a symmetric way, while in NK they are treated differently. In the formula for Lorentz transformations, the divisors $/\sqrt{1-v^2}$ represent time-dilation and length-contraction, while "$t - vx$" in place of "t" in the first part represents "getting out of synchronism".

By a *space-isometry* (over $\mathfrak{Q}$) we understand a Euclidean isometry (i.e. a mapping that preserves Euclidean distance between space-time locations) which takes the time-axis to a line parallel to the time-axis. These are *affine* mappings, i.e. linear mappings composed with translations.

THEOREM 11.10 (DESCRIPTION OF WORLDVIEW TRANSFORMATIONS OF **Specrel**) *Assume $n \geq 3$, let $\mathfrak{Q} = \langle Q, +, *, \leq\rangle$ be a quadratic ordered field and let $f : Q^n \to Q^n$. The following are equivalent.*

(i) *f is a worldview transformation in a model of* **Specrel** *with field-reduct $\mathfrak{Q}$.*

(ii) *$f = \sigma \circ \lambda \circ \sigma'$ for some Lorentz transformation λ and space-isometries σ, σ' (over $\mathfrak{Q}$).*

(iii) *f is a bijection and preserves relativistic distance, i.e.* $(\forall p, q \in \mathsf{Q}^n)\boldsymbol{\mu}(p,q) = \boldsymbol{\mu}(f(p), f(q))$.

On the proof (i)⇒(ii): By Theorem 11.4(i) we know that w_{mk} is like a Lorentz transformation on the "*plane of motion*", i.e. on the vertical plane containing $\mathsf{wline}_m(k)$, and it takes the subspace of Q^n orthogonal to this plane to itself, by Theorem 11.4(ii). AxSim then states that w_{mk} is a Euclidean isometry on this orthogonal subspace. The proof of (i)⇒(ii) from here on is not difficult. (ii)⇒(iii): A possibility for proving this is that one checks by a computation that both space-isometries and Lorentz transformations preserve relativistic distance. However, we would like to provide more insight here concerning (ii)⇒(iii). Namely, showing that Lorentz transformations preserve lines of slope 1 (i.e. that they preserve $\boldsymbol{\mu}(p,q) = 0$) is the most important step in proving that **Specrel** is consistent. Fig. 11.18 illustrates a non-computational, geometric proof for this crucial part of the proof. (iii)⇒(i): If f preserves $\boldsymbol{\mu}$, then f preserves lines of slope 1, preserves lines of slope < 1, and also "it satisfies

AxSim". The rest follows from the construction we give after Theorem 11.11.
QED

By a *space-dilation* (over $\mathfrak{Q}$) we understand a Euclidean dilation (i.e. a mapping that "dilates" Euclidean distances between space-time locations with a given ratio $r \in Q$) and takes the time-axis to a line parallel to the time-axis. These are affine mappings. By a *field-automorphism-induced mapping* over $\mathfrak{Q}$ we understand the natural extension of an automorphism of $\mathfrak{Q}$ to Q^n. These are not necessarily affine mappings, but they are *collineations*, i.e. they take straight lines to straight lines.

THEOREM 11.11 (DESCRIPTION OF WORLDVIEW TRANSFORMATIONS OF **Specrel**$_0$) *Assume $n \geq 3$, let $\langle Q, +, *, \leq \rangle$ be a quadratic ordered field and let $f : Q^n \to Q^n$. The following are equivalent.*

(i) *f is a worldview transformation in a model of* **Specrel**$_0$.

(ii) *$f = \delta \circ \lambda \circ \delta' \circ \alpha$ for some Lorentz transformation λ, space-dilations δ, δ', and field-automorphism-induced mapping α.*

(iii) *f is a bijection and preserves relativistic distance 0, i.e.*
$(\forall p, q \in \mathsf{Q}^n)[\boldsymbol{\mu}(p, q) = 0 \quad \textit{iff} \quad \boldsymbol{\mu}(f(p), f(q)) = 0]$.

On the proof (ii)$\Leftrightarrow$(iii) is a variant of the Alexandrov-Zeeman theorem, see e.g. Goldblatt, 1987, App. 2. (i)$\Rightarrow$(iii) follows from AxPh, and (ii)$\Rightarrow$(i) follows from the construction we give soon, because (ii) implies that f preserves lines of slope 1 and lines of slope < 1. QED

Remark: We could have proved Theorems 11.4–11.7 by first deriving the properties of the worldview transformations as in Theorems 11.10, 11.11, and then deriving the paradigmatic effects (i)–(iii) from their properties. We think that deriving the predictions of relativity theory directly from the axioms is more illuminating. For the student, Lorentz transformations appear as non-observation oriented theoretical constructions not explaining why we are doing what we are doing. In our approach, stating the axioms in **Specrel** and then deriving the three paradigmatic effects motivate the introduction of Lorentz transformations. In some sense, **Specrel** can be considered as an "implicit definition" of the Lorentz transformations.

We now turn to an exhaustive description of all models of **Specrel**$_0$ and **Specrel**.

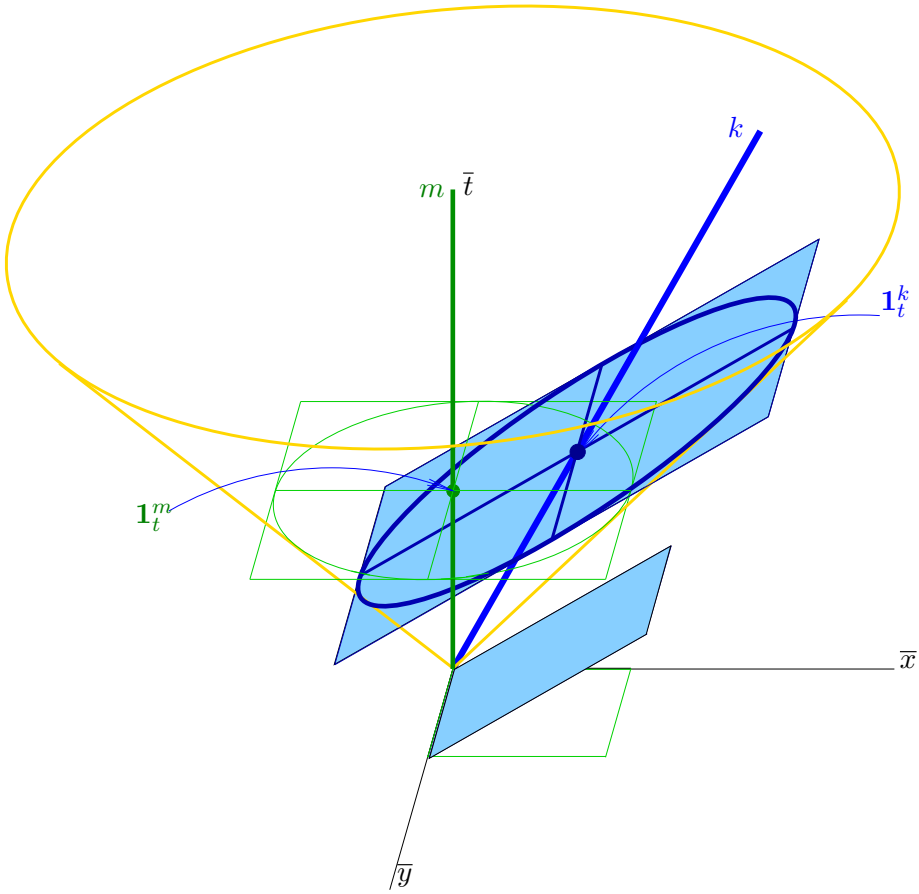


Figure 11.18. The Light Axiom states that if we switch on a light source for a moment, we will observe a light-sphere expanding away from us with the speed of light, and that we are all the time in the center of this light-sphere. Assume that observers m, k are present in the event of switching on the light source and that they are moving relative to each other. Then both observers m and k have to observe that they are in the center of the photon-sphere! How is this possible? This figure illustrates how. Let $\overline{x}$ be in the direction of movement of k in m's worldview, and let $\overline{y}$ be any spatial direction orthogonal to $\overline{x}$. The expanding light-sphere in space-time when concentrating on the 3-dimensional subspace determined by $\overline{t}, \overline{x}, \overline{y}$ is a cone. k's plane of simultaneity is tilted just so that k is in the center of the ellipse that is the intersection of this plane with the light-cone. For this, the "long axis" of the ellipse is tilted just the amount that it is symmetric to the worldline of k (w.r.t. a photon-line as in Fig. 11.17), and the "small axis" of the ellipse is parallel with $\overline{y}$. This implies "k's clocks getting out of synchronism" (Theorem 11.4). The time-unit $\mathbf{1}_t^k$ of k is on k's worldline exactly so "high" that the length of the "short axis" of the ellipse is 1, when AxSim is true (and arbitrary otherwise). This implies that k's time flows slowly as seen by m when AxSim is true. (Paradigmatic effects in Theorem 11.6!) The other space-unit $\mathbf{1}_x^k$ of k is chosen so that the length of the "long axis" of the ellipse counts as 1, too. With this choice of the units of measurement, k sees the ellipse as a circle, hence k thinks that he is in the center of the light-cone. There is enough room for everyone in the center of the expanding lightsphere!

Given an arbitrary quadratic ordered field $\mathfrak{Q}$, let PLines and TLines denote the set of all straight lines in Q^n of slope 1 and of slope < 1, respectively. Let WT denote the set of all transformations of Q^n which preserve both PLines and TLines. Then WT forms a group, and Theorem 11.11 describes the members of WT.

Consider $\mathfrak{M} = \langle \mathfrak{Q};\ B, Ob, Ph;\ W\rangle \models \mathbf{Specrel}_0$ in which $Ob \neq \emptyset$. Then Ph and Ob are disjoint, by Theorem 11.7(i). We let

$$B_1 := B - (Ph \cup Ob).$$

In the discussion below we will see that (after having chosen $\mathfrak{Q}$), the "heart" of a $\mathbf{Specrel}_0$ model is a subgroup $\mathsf{WT}_0 \subseteq \mathsf{WT}$ such that $\{f^{-1}[\bar{t}\,] : f \in \mathsf{WT}_0\} = \mathsf{TLines}$. Here $f[H] := \{f(a) : a \in H\}$ for any function f and subset H of the domain $\mathsf{Dom}(f)$ of f, as usual. Having chosen the heart WT_0 of our model, we still have to decorate it with "observer names" (Ob), "photon names" (Ph), and with extra bodies B_1 not necessarily in $\mathsf{Ob} \cup \mathsf{Ph}$. This decorating or labelling gives rise to an extra plurality of (nonisomorphic) possible models for $\mathbf{Specrel}_0$ in addition to the possible choices of WT_0. (These choices will be formally specified in items (i)-(v) below.) Below we present the details giving precise meanings to what we understand by the above, e.g. by "labelling", "heart" etc. The reader not interested in the details might skim over them just to have an impression and continue serious reading with Corollary 11.12.

Let us return to our $\mathfrak{M} = \langle \mathfrak{Q};\ B, Ob, Ph;\ W\rangle \models \mathbf{Specrel}_0$. For any observer $m \in Ob$ let us define the "*worldview of* m" as a structure

$$\mathsf{w}_m := \langle Q^n, \mathsf{wline}_m(b) : b \in B\rangle$$

where Q^n is the carrier set and $\mathsf{wline}_m(b)$ is a one-place relation (or unary predicate) with relation symbol b (denoting this subset of Q^n) for each $b \in B$ (recall that $\mathsf{wline}_m(b) \subseteq \mathsf{Q}^n$, and cf. Fig. 11.1 on p.611). The set of all worldviews contains exactly the information content of $W \cap Ob \times B \times Q^n$. Let $W_1 := W - Ob \times B \times Q^n$. When we want to define a model of $\mathbf{Specrel}_0$, we have to define the 6-place relation W. Instead of defining W directly, often it is easier to define the set of worldviews along with W_1.

All these worldviews are isomorphic with each other, actually the worldview transformations are isomorphisms between the worldviews, i.e. $\mathsf{w}_{mk} : \mathsf{w}_m \to \mathsf{w}_k$ is an isomorphism for any $m, k \in Ob$, by the definition of w_{mk}. What do the worldviews w_m look like? The carrier set is Q^n, the photons are distributed surjectively on the PLines by AxPh; the observers are distributed surjectively on the TLines by AxLine, AxThEx, Theorem 11.7; $\mathsf{wline}_m(m) = \bar{t}$ by AxSelf. Instead of specifying all the worldviews one-by-one, we can specify one "Platonic" (or generic) worldview

$$\mathfrak{P} := \langle Q^n, \pi(b) : b \in B\rangle, \qquad \text{where}$$

$\pi \restriction Ph : Ph \twoheadrightarrow \mathsf{PLines}$ is surjective,

$\pi \restriction Ob : Ob \twoheadrightarrow \mathsf{TLines}$ is surjective, and

$\pi \restriction B_1 : B_1 \to \{Y : Y \subseteq Q^n\}$,

and then for each observer $m \in Ob$ we can specify an element of WT which describes how m realizes this Platonic worldview, i.e. we specify a function

$w : Ob \to \mathsf{WT}$ such that $w(m)[\pi(m)] = \bar{t}$;

and then we define w_m as the image of $\mathfrak{P}$ by the function $w(m)$. The information content of $\pi \restriction Ob : Ob \twoheadrightarrow \mathsf{TLines}$ can be recovered from this last function w (by $\pi(m) = w(m)^{-1}[\bar{t}]$), so we may skip specifying $\pi \restriction Ob$.

With the above intuition in mind, we can construct a model of **Specrel**$_0$ by specifying a quintuple $\langle \mathfrak{Q}, w, \pi, \beta, W_1 \rangle$ with the following properties:

(i) $\mathfrak{Q}$ is a quadratic ordered field,

(ii) w, π, β are functions with disjoint domains Ob, Ph, B_1 respectively,

(iii) $w : Ob \to \mathsf{WT}$ is such that $\{w(m)^{-1}[\bar{t}] : m \in Ob\} = \mathsf{TLines}$,

(iv) $\pi : Ph \twoheadrightarrow \mathsf{PLines}$ is surjective,

(v) $\beta : B_1 \to \{Y : Y \subseteq Q^n\}$, and $W_1 \subseteq (B_1 \cup Ph) \times (B_1 \cup Ph \cup Ob) \times Q^n$.

Let us call a quintuple satisfying the above conditions a *pre-model*. From any pre-model $\langle \mathfrak{Q}, w, \pi, \beta, W_1 \rangle$ we can construct a model of **Specrel**$_0$ by defining

$\mathfrak{M}(\mathfrak{Q}, w, \pi, \beta, W_1) := \langle \mathfrak{Q};\ B, Ob, Ph;\ W \rangle$ where

$Ob := \mathsf{Dom}(w)$, $Ph := \mathsf{Dom}(\pi)$, $B := \mathsf{Dom}(\beta) \cup Ob \cup Ph$; and W is defined the natural way

$$W := \{\langle m, b, p \rangle : p \in w(m)[\beta(b)], m \in Ob\} \cup \{\langle m, \mathsf{ph}, p \rangle : p \in w(m)[\pi(\mathsf{ph})], m \in Ob\} \cup \{\langle m, k, p \rangle : p \in w(m)[w(k)^{-1}[\bar{t}]], m \in Ob\} \cup W_1.$$

It can be checked that all models constructed from pre-models are models of **Specrel**$_0$; and conversely, all models of **Specrel**$_0$ with nonempty observer-part arise this way from pre-models.

We described the models of **Specrel**$_0$ with $\mathsf{Ob} \neq \emptyset$. The description when $\mathsf{Ob} = \emptyset$ is easy. The description of the models of **Specrel** is exactly like above with the only change that in place of WT we use its subset WT^+ characterized in Theorem 11.10.

COROLLARY 11.12 (CONSISTENCY) **Specrel** *is a consistent theory, i.e. for no formula φ can both φ and its negation $\neg\varphi$ be derived from* **Specrel**. *Moreover,* **Specrel** $+ (\mathsf{Ob} \neq \emptyset)$ *is also consistent.*

Proof To construct a model of **Specrel** + ($\mathsf{Ob} \neq \emptyset$) we have to show that there exists at least one pre-model $\langle \mathfrak{Q}, w, \pi, \beta, W_1 \rangle$ with $w : Ob \to \mathsf{WT}^+$. Of the conditions (i)–(v) in the definition of a pre-model, only condition (iii) is not trivial to satisfy. However, Theorem 11.10(ii) shows that for all $\ell \in \mathsf{TLines}$ there is $w \in \mathsf{WT}^+$ which takes $\bar{t}$ to ℓ, and we are done. QED

Corollary 11.12 above completes justification of the move NK $\mapsto$ NK$^-$. It shows that by this move, we indeed got rid of all contradictions between NK and the Light Axiom.

Having a description of all models of **Specrel** and **Specrel**$_0$ at hand makes it easy to see which statements follow from **Specrel** and **Specrel**$_0$ and which do not. As an example we include the following.

COROLLARY 11.13 *Let* $n \geq 3$. **Specrel** $\vdash (\forall m, k \in \mathsf{Ob}) v_m(k) = v_k(m)$ *while* **Specrel**$_0$ $\nvdash (\forall m, k \in \mathsf{Ob}) v_m(k) = v_k(m)$.

Hint for proof Field-automorphism-induced mappings α can occur in worldview-transformations in models of **Specrel**$_0$, but not in models of **Specrel**. QED

THEOREM 11.14 (INDEPENDENCE OF THE AXIOMS) *Assume* $n \geq 3$.

(i) (**Specrel**$_0$ $- \{\mathsf{AxLine}, \mathsf{AxThEx}\}$) $\vdash \mathsf{AxLine}$.

(ii) **Specrel** $- \{\mathsf{AxLine}\}$ *is an* independent *axiom system, i.e.* (**Specrel** $- \{\mathsf{Ax}\}$) $\nvdash \mathsf{Ax}$ *for any element* Ax *in* **Specrel** *different from* AxLine.

(iii) *Every model of* **Specrel**$_0$ $- \{\mathsf{AxThEx}\}$ *can be extended to a model of* **Specrel**$_0$. *Hence if formula* η *is universally quantified in the sort* B *and* **Specrel**$_0$ $\vdash \eta$, *then* (**Specrel**$_0$ $- \{\mathsf{AxLine}, \mathsf{AxThEx}\}$) $\vdash \eta$. *The same holds for* **Specrel** *in place of* **Specrel**$_0$. QED

By Theorem 11.14(iii) above, all our paradigmatic effects can be proved in the more economical fragment **Specrel** $- \{\mathsf{AxLine}, \mathsf{AxThEx}\}$ of **Specrel**. (This is so because the paradigmatic effects can be reformulated as sentences universally quantified in sort B, by using Theorem 11.2.) On the other hand, Theorems 11.4, 11.6 do not hold if we omit any one of the axioms of **Specrel**$_0$ $- \{\mathsf{AxLine}, \mathsf{AxThEx}\}$, and Theorem 11.6 does not hold if we omit AxSim. Such investigations of economy asking which axioms are needed for proving what theorem are called "reverse relativity theory" motivated by the highly successful branch of mathematics called "reverse mathematics" and is pursued in Andréka et al., 2002. We will return to this important subject in Sec. 2.7.

Specrel has many non-elementarily equivalent models over any quadratic ordered field. We show that **Specrel** can be extended to a theory **Specrel** $\cup$ **Comp** which is categorical over any quadratic ordered field, it can be extended to a

complete and decidable theory, and **Specrel** can also be extended to a hereditarily undecidable theory. Both extensions are natural. (Cf. Theorem 11.15 below.)

Recall that from any pre-model $\langle \mathfrak{Q}, w, \pi, B_1, W_1 \rangle$ we can construct a model of $\mathbf{Specrel}_0$. The most natural pre-models for **Specrel** are $\langle \mathfrak{Q}, \mathsf{Id} \restriction \mathsf{WT}^+, \mathsf{Id} \restriction \mathsf{PLines}, \emptyset, \emptyset \rangle$. We will call the models constructed from these *standard models* for **Specrel**. Thus, in the standard models we include all the possible kinds of observers, but otherwise we are as "economic" as possible. There is exactly one standard model of dimension n over any quadratic ordered field $\mathfrak{Q}$. We are going to give a complete axiom system for these standard models.

$\mathsf{AxCoord}$ $(\forall m \in \mathsf{Ob})(\forall \text{ space-isometry } S \text{ of } \mathsf{Q}^n)(\exists k \in \mathsf{Ob})\mathsf{w}_{mk} = S.$

AxExt^{ob} $(\forall m, k \in \mathsf{Ob})(\mathsf{w}_{mk} = \mathsf{Id} \to m = k).$

AxExt^{ph} $(\forall \mathsf{ph}, \mathsf{ph}' \in \mathsf{Ph})(\forall m \in \mathsf{Ob})(\mathsf{wline}_m(\mathsf{ph}) = \mathsf{wline}_m(\mathsf{ph}') \to \mathsf{ph} = \mathsf{ph}').$

$\mathsf{AxNobody}$ $\mathsf{B} = \mathsf{Ob} \cup \mathsf{Ph}$ and $\mathsf{W} \subseteq \mathsf{Ob} \times \mathsf{B} \times \mathsf{Q}^{n-2}$.

$\mathbf{Comp} := \{\mathsf{AxCoord}, \mathsf{AxExt}^{ob}, \mathsf{AxExt}^{ph}, \mathsf{AxNobody}\}.$

The above are natural axioms which hold in all standard models of **Specrel**. $\mathsf{AxCoord}$ expresses that each observer can "re-coordinatize" his worldview with a space-isometry. There is a quantifier ranging over space-isometries in this formula. Nevertheless, this axiom can be expressed with a first-order logic formula because space-isometries are affine mappings and hence can be "coded" with the images of the n unit-vectors $\mathbf{1}_i$. The next two axioms in **Comp** say, intuitively, that of each kind of observers and photons we have only one copy (or, in other words, according to Leibniz's principle, if we cannot distinguish two observers or photons with some observable properties expressible in our language, then we treat them as equal). Hence we call them extensionality principles, whence the abbreviation AxExt. The example of AxExt^{ph} reveals that here we consider only space-time-oriented properties of photons, hence two photons of different "color" but same worldline are not distinguished in the theory **Comp**. AxExt^{ob} also expresses that we really identify observers with coordinate systems. These axioms are natural to assume, we did not include them in **Specrel** because these "simplifying axioms" are not needed for proving the theorems. The last axiom in **Comp** says that every body is an inertial observer or photon. This is a real restriction that we usually do not want to make when we use AxLine. The main reason is that we can treat accelerated observers in **Specrel** if we do not make this restriction $\mathsf{AxNobody}$, see Sec. 3.1. Treating accelerated observers in **Specrel** is important for the transition from special relativity theory to general relativity theory, as we shall see. $\mathsf{AxNobody}$ excludes accelerated bodies. So we do "pay a physical price" for assuming $\mathsf{AxNobody}$.

By a *theory* we understand an arbitrary set of first-order logic formulas (i.e. we will not assume that a theory contains all its semantical consequences). We call a theory Th *decidable* (or *undecidable* respectively) if the set of all first-order logic semantical consequences of Th is decidable (or undecidable respectively). We call Th *complete* if it implies either φ or $\neg\varphi$ for each first-order logic formula φ without free variables (of its language). It is known that the theory of quadratic ordered fields is undecidable. A quadratic ordered field is called *real-closed* if every polynomial of odd degree has zero as a value. This last requirement can be expressed with the infinite set $RC := \{\phi_{2n+1} : n \in \omega\}$ of first-order logic formulas, where ϕ_n denotes the following sentence

$$\forall x_0 \ldots \forall x_n \exists y (x_n \neq 0 \rightarrow x_0 + x_1 \cdot y + \cdots + x_n \cdot y^n = 0).$$

Tarski proved that the theory of real-closed fields is complete and decidable (cf., e.g. Hodges, 1993, Theorem 2.7.2, p. 67, p. 92). The above suggests that if we want to obtain interesting and relevant decidability-theoretic results, then we have to concentrate on real-closed fields; or at least include a decidable theory of field-axioms into our theories.

THEOREM 11.15 *Let* $n \geq 3$.

(i) *The models of* **Specrel** $\cup$ **Comp** *are exactly the models isomorphic to standard ones.*

(ii) **Specrel** $\cup$ **Comp** $\cup\, TF$ *is complete and decidable, for any complete, decidable theory* TF *of quadratic ordered fields.*

(iii) **Specrel** $\cup$ (**Comp** $-$ $\{\mathsf{Ax}\}$) $\cup\, RC$ *can be extended to a hereditarily undecidable theory* Th *for any* $\mathsf{Ax} \in$ **Comp**, *in the sense that no consistent extension of* Th *is decidable.*

For proof of Theorem 11.15 and for related results we refer to Andréka et al., 1999, Sec. 7, Andréka et al., 2004, Sec. 7.

In the standard models of **Specrel** there is no orientation for time, "reversing time" is an automorphism of these models. In relativity theory, both special and general, we sometimes use the fact that "time has a direction", i.e. that $\mathsf{Ax}\uparrow$ below is assumed.

$\mathsf{Ax}\uparrow$ $(\forall m, k \in \mathsf{Ob})(\forall p, q \in \bar{t})[\,(v_m(k) < 1 \wedge p_t \leq q_t) \Rightarrow \mathsf{w}_{km}(p)_t \leq \mathsf{w}_{km}(q)_t\,]\,.$

Axiom $\mathsf{Ax}\uparrow$ expresses that every observer sees the time of another slowly moving observer "flow forwards", i.e. m sees k's clocks ticking "forwards" and not "backwards". All our theorems so far are true for **Specrel** $+$ $\mathsf{Ax}\uparrow$ in place of **Specrel** with minor modifications.

2.6 Observer-independent geometries in relativity theory; duality and definability theory of logic

According to the approach taken so far, each observer observes the world through the "looking-glass" or "spectacles" of his own coordinate system. The question comes up: Is there an observer-independent, "absolute" reality which the individual observers observe through their respective coordinate systems, or is the set of worldviews of the different observers just an ad-hoc collection of subjective personal views? (The philosophy of subjective idealism contra the assumption of the existence of an objective external world.) We will see that relativistic space-time (or relativistic geometry) provides such an observer-independent reality. Namely, in the present subsection we show that the observers (and their coordinate systems) can be defined from relativistic distance $\boldsymbol{\mu}$ as defined at the end of Sec. 2.5, in models of **Catrel** := **Specrel** $\cup$ **Comp**. If we regard the worldviews of the various observers as "subjective" (in some sense), then $\boldsymbol{\mu}$ is "objective" in the sense that $\boldsymbol{\mu}$ is the same for all observers, in **Specrel**. Thus relativistic distance $\boldsymbol{\mu}$ provides such an observer-independent, absolute reality. This statement will be made more tangible in the definition of the observer-independent geometry $\mathsf{Mg}(\mathfrak{M})$ associated to **Specrel** models $\mathfrak{M}$ below.

We already saw that in models of $\mathbf{Specrel}_0$, all observers observe the same events. Let Events denote the set of events observed by some (or equivalently by each) observer,

$$\mathsf{Events} := \{\mathsf{ev}_m(p) : m \in \mathsf{Ob},\ p \in \mathsf{Q}^n\}.$$

In models of **Specrel**, we can define relativistic distance $\boldsymbol{\mu}$ of events, by Theorem 11.8, as

$$\boldsymbol{\mu}(e, e') := \boldsymbol{\mu}(\mathsf{loc}_m(e), \mathsf{loc}_m(e')), \quad \text{for any observer } m \in \mathsf{Ob}.$$

Relativistic distance of events is a function $\boldsymbol{\mu} : \mathsf{Events} \times \mathsf{Events} \to \mathsf{Q}$. Given $\mathfrak{M} \models$ **Specrel** we define its *metric-geometry* (or *Minkowski geometry*) $\mathsf{Mg}(\mathfrak{M})$ as a two-sorted structure as follows:

$$\mathsf{Mg}(\mathfrak{M}) := \langle \mathsf{Events}, \boldsymbol{\mu};\ \mathsf{Q}, 1\rangle.$$

$\mathsf{Mg}(\mathfrak{M})$ is also referred to as the *space-time* of $\mathfrak{M}$. The two sorts of $\mathsf{Mg}(\mathfrak{M})$ are Events and Q; $\boldsymbol{\mu}$ is a function of sort $\mathsf{Events} \times \mathsf{Events} \to \mathsf{Q}$ and 1 is a constant of sort Q. We want to state a strong equivalence between $\mathsf{Mg}(\mathfrak{M})$ and $\mathfrak{M}$. These two structures have different vocabularies (or signatures, or languages).

The part of logic that connects structures and theories on different vocabularies is called *definability theory*. The strongest kind of connection between two theories is *definitional equivalence* of theories. When two theories are definitionally equivalent, we say that they are lexicographical variants of the same

theory, the only difference being that they use different concepts of the theory as basic ones. We will need definitional equivalence of first-order logic theories on vocabularies that have different *sorts* (or universes), hence the definitionally equivalent models will have different kinds of universes. This amounts to defining new "entities" in a model, not only new relations or functions on already existing entities as in "standard" one-sorted definability theory of first-order logic. Definability of new sorts is important in the kind of definability that arises in relativity theory; therefore we worked out such a definability theory in Andréka et al., 2002, Sec. 6.3 and Madarász, 2002, Sec. 4.3. Below we recall the elements of definability theory that we are going to use.

Let L be a vocabulary, possibly many-sorted, and let L' be an *expansion of* L, i.e. L' may contain new sorts, and new relation and function symbols. The L*-reduct* of a structure $\mathfrak{M}'$ on vocabulary L' is the obvious thing (we "forget" the interpretations of symbols not in L). Let Th and Th' be theories on vocabularies L and L', respectively. We say that Th' is a *definitional expansion* of Th iff the following (i)-(ii) hold:

(i) The models of Th are exactly the L-reducts of models of Th'.

(ii) For any two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$ of Th' with the same L-reduct there is a unique isomorphism between $\mathfrak{M}_1$ and $\mathfrak{M}_2$ that is the identity on this common reduct.

Let Th_1 and Th_2 be arbitrary theories (possibly on completely different vocabularies). We say that Th_1 and Th_2 are *definitionally equivalent* when they have a joint definitional expansion Th_3. (More precisely, definitional equivalence is the transitive closure of the notion just defined. In this subsection we will not need to take transitive closure.)

Intuitive explanation for definitional equivalence of theories: By "Th' is a definitional expansion of Th" we mean that each sort, relation and function in the vocabulary of Th' that is not present in the vocabulary of Th is actually defined in Th' in the following sense. Properties (i)-(ii) above express that we consider a new relation (i.e. one in L' but not in L) on an existing sort as defined (in Th') if it can be "put" on every model of Th in a unique way so that it satisfies Th', and we consider a new sort (and relations on it) as defined if it can be added to each model of Th in a unique way, up to a unique isomorphism.

In definability theory of logic, there are a semantical and a syntactical approach to definability, and of course the interesting thing is to state their equivalence (this is Beth's theorem in the usual definability theory of first-order logic). We presented here the notions of the semantic approach; when Th' is a definitional expansion of Th we can say that Th' is an "implicit, or semantical definition" of the symbols not occurring in Th. In Andréka et al., 2002, Sec. 6.3 and Madarász, 2002, Sec. 4.3 we worked out the "syntactical" counterpart of

this definability, i.e. we gave concrete prescriptions for what an "explicit, syntactical definition" of a new element of the vocabulary can look like. Then we proved the analogue of Beth's theorem (stating that a new element of the vocabulary has an implicit definition exactly when it has an explicit definition). When two theories are definitionally equivalent, there is a computable meaning-preserving translation function between their languages (see Madarász, 2002, 4.3.27 and 4.3.29). For more on definability theory we refer to e.g. Makkai, 1993.

We now proceed to state definitional equivalence between theories occurring in relativity theory. From now on, in the present subsection, we assume $n \geq 3$.

The formula $\boldsymbol{\mu}(p,q) = (p_1 - q_1)^2 - (p_2 - q_2)^2 - \cdots - (p_n - q_n)^2$ is referred to as the (squared) Minkowski metric and the structure $\langle \mathsf{Q}^n, \boldsymbol{\mu};\ \mathsf{Q}, 1\rangle$ is the (metric) Minkowski geometry over $\mathfrak{Q}$. For a class K of structures, IK denotes the class of all structures isomorphic to elements of K.

THEOREM 11.16 (DEFINITIONAL EQUIVALENCE BETWEEN METRIC-GEOMETRIES AND **Catrel** MODELS) *Assume* $n \geq 3$.

(i) **Catrel** *is definitionally equivalent to the first-order logic theory of its metric-geometries, i.e.* **Catrel** *and* Th_m *in (ii) below are definitionally equivalent.*

(ii) $\mathsf{I}\{\langle \mathsf{Events}, \boldsymbol{\mu};\ \mathsf{Q}, 1\rangle : \mathfrak{M} \models \mathbf{Catrel}\} =: \mathsf{MG}$ *is axiomatizable by finitely many formulas, i.e. there is a finite axiom system* Th_m *such that* MG *is the class of all models of* Th_m.

(iii) $\mathsf{I}\{\langle \mathsf{Q}^n, \boldsymbol{\mu};\ \mathsf{Q}, 1\rangle : \mathsf{Q}$ *is a quadratic ofield*$\} = \mathsf{MG}$. *I.e., the class of Minkowski geometries (over quadratic ordered fields) coincides with the class of metric-geometric models of* **Catrel**.

Outline of proof **(i):** We define a joint definitional expansion Th_3. The vocabulary of Th_3 contains all the symbols occurring either in **Specrel** or in $\mathsf{Mg}(\mathfrak{M})$, plus one new $n+2$-place relation symbol J of type $\mathsf{B} \times \mathsf{Events}^{n+1}$. Let $\mathfrak{M} = \langle \mathsf{Q}, +, *, \leq;\ \mathsf{B}, \mathsf{Ob}, \mathsf{Ph};\ \mathsf{W}\rangle \models \mathbf{Specrel}_0$ be given, and define

$$J_{\mathfrak{M}} := \{\langle m, \mathsf{ev}_m(\overline{0}), \mathsf{ev}_m(\mathbf{1}_t), \ldots, \mathsf{ev}_m(\mathbf{1}_n)\rangle : m \in \mathsf{Ob}\} \cup \{\langle \mathsf{ph}, e_1, \ldots, e_{n+1}\rangle : \mathsf{ph} \in \mathsf{Ph}, e_1, \ldots, e_{n+1} \in \mathsf{Events}, \mathsf{ph} \in e_1, \ldots, \mathsf{ph} \in e_{n+1}\}, \quad \text{and the expansion of } \mathfrak{M}$$

$$F(\mathfrak{M}) := \langle \mathsf{Q}, +, *, \leq, 0, 1;\ \mathsf{B}, \mathsf{Ob}, \mathsf{Ph};\ \mathsf{Events}, \boldsymbol{\mu};\ \mathsf{W}, J_{\mathfrak{M}}\rangle,$$

Th_3 is the set of all formulas valid in $\{F(\mathfrak{M}) : \mathfrak{M} \models \mathbf{Catrel}\}$.

Clearly, $J_{\mathfrak{M}}$ gives an "interpretation" of observers and photons in Events, and so makes a connection between the two "alien" sorts B and Events. The following are the main ideas in showing that Th_3 is a definitional expansion of Th_m (the theory of metric-geometries of **Catrel**). We have $\mathsf{Events}, \boldsymbol{\mu}, \mathsf{Q}, 1$ at our disposal and we have to "define" (or recover) $+, *, \leq, 0, \mathsf{B}, \mathsf{Ob}, \mathsf{Ph}, \mathsf{W}, J$. First we define $0 := \boldsymbol{\mu}(e, e)$, then we define "lightlike collinearity" on Events by using $\boldsymbol{\mu}$ as follows: e_1, e_2, e_3 are *lightlike collinear* iff $\boldsymbol{\mu}(e_i, e_j) = 0$ for all $i, j = 1, 2, 3$. From lightlike collinearity then we define usual *collinearity* as in the proof of the Alexandrov-Zeeman theorem in Goldblatt, 1987, App. 2, or in Andréka et al., 1999, Andréka et al., 2004. For the idea of this part of the proof see Fig. 11.19. A proof for (a generalization of) the Alexandrov-Zeeman theorem using a different, elegant idea is in Horváth, 2005. A definability-theoretic analysis of the Alexandrov-Zeeman theorem in an axiomatic setting can be found in Pambuccian, 2006. From collinearity and $0, 1, \boldsymbol{\mu}$ we define the *field-operations* $+, *$ by using Hilbert's coordinatization technique (see e.g. Goldblatt, 1987, pp. 23-27 or Andréka et al., 2002, Sec. 6.5.2). Since the original field was quadratic and ordered, we can recover the *ordering* $\leq$, too. From collinearity and $\boldsymbol{\mu}$ we can define the so-called relativistic (or Minkowski) *orthogonality* relation (see Fig. 11.22), and from $\boldsymbol{\mu}$ again then we can define the $n+1$-tuples of events $\langle e_0, e_1, \ldots, e_n \rangle$ that correspond exactly to $\langle \mathsf{ev}_m(\overline{0}), \mathsf{ev}_m(\mathbf{1}_t), \ldots, \mathsf{ev}_m(\mathbf{1}_n) \rangle$ for some observer m by requiring that $\boldsymbol{\mu}(e_0, e_i) = 1$ and e_0, e_i is orthogonal to e_0, e_j for all $i, j = 1, \ldots, n$, $i \neq j$. We can use these to define $\mathsf{Ob}, \mathsf{Ph}, \mathsf{B}, J$ and W. In the above we made use of the fact that all the constructions can be tracked with first-order logic formulas. Showing that Th_3 is a definitional expansion of **Catrel** is the easier direction, for a proof see Madarász, 2002, p. 241. **(ii):** In the above construction, we defined the operations of **Catrel** by using $\boldsymbol{\mu}, 1$, thus we can express the finitely many axioms defining **Catrel** by using $\boldsymbol{\mu}, 1$ and we are done. **(iii):** The functions ev_m and loc_m define isomorphisms between the structures $\langle \mathsf{Q}^n, \boldsymbol{\mu};\ \mathsf{Q}, 1 \rangle$ and $\langle \mathsf{Events}, \boldsymbol{\mu};\ \mathsf{Q}, 1 \rangle$. QED

From the proof of Theorem 11.16 we actually can construct a finite theory Th_m axiomatizing the metric-geometries MG. This Th_m, however, is complicated and not really illuminating. It would be nice to find a streamlined, finite axiom system Th axiomatizing MG which contains few and easy-to-understand, illuminating axioms about $\boldsymbol{\mu}, 1$.

We called a property absolute if every observer "sees" it the same way. This "absolute" means also "observer-independent" or "coordinate-independent". Theorem 11.8 states that relativistic distance $\boldsymbol{\mu}$ is such an absolute property. Clearly, every formula expressible by the use of $\boldsymbol{\mu}, 1$ is absolute, too. The corollary below says that these are all the absolute coordinate properties of events, in **Specrel**. By a coordinate property of events we understand a property expressible in terms of the coordinates $\mathsf{loc}_m(e)$ of events e; more concretely

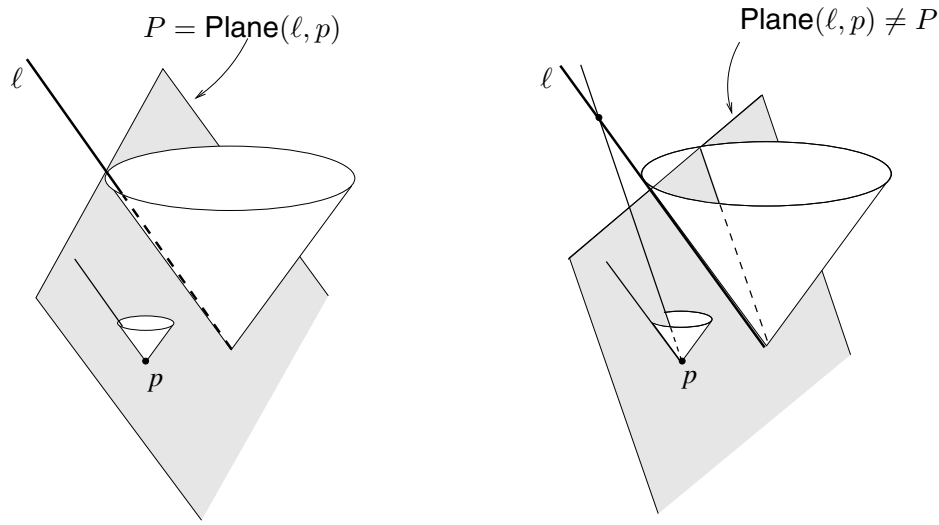


Figure 11.19. Collinearity can be defined from lightlike collinearity, as follows. Assume $n = 3$. Given a lightline ℓ, the plane P tangent to the light-cone and containing ℓ is the set of those points p through which no lightline intersecting ℓ goes. The reason for this is illustrated in the two parts of the figure. Then we get all spacelike lines as intersections of tangent planes. Then each timelike plane can be defined by a pair of intersecting lightlines and the spacelike lines connecting them; timelike lines then are the "new" intersections of timelike planes. In the above, spacelike, timelike lines, and lightlines are straight lines that lie outside, inside, and on the light-cone, respectively. A plane is timelike if it contains a timelike line. The case $n > 3$ is similar.

by a coordinate property of events $e_1, \ldots, e_r$ we understand a formula either in the form $(\forall m \in \mathsf{Ob})\psi(\mathsf{loc}_m(e_1), \ldots, \mathsf{loc}_m(e_r))$ or in the form $(\exists m \in \mathsf{Ob})\psi(\mathsf{loc}_m(e_1), \ldots, \mathsf{loc}_m(e_r))$ where ψ is a formula in the vocabulary of the field-reduct $\langle Q, +, *, \leq\rangle^n$.

COROLLARY 11.17 *Every relation definable (in FOL) on* Events *in a model of* **Catrel** *can be defined from* $\boldsymbol{\mu}$ *and* 1. *Every coordinate-property of events in a model of* **Specrel** *can be defined from* $\boldsymbol{\mu}$ *and* 1.

Theorem 11.16 can be interpreted as saying that metric-geometries are the observer-independent, "absolute realities" corresponding to models of **Specrel**. If we abstract from the concrete values of the metric-properties in the metric-geometries, we get the so-called causal-geometries, to be defined below. These correspond to the "absolute realities" of models of $\mathbf{Specrel}_0$. We are going to elaborate these ideas.

Let us call events e, e' *causally separated*, in symbols $e \sim_c e'$, iff there is either an observer or a photon that participates both in e and in e', i.e. iff $e \cap e' \cap$

$(\mathsf{Ob} \cup \mathsf{Ph}) \neq \emptyset$. We call two space-time locations $p, q \in \mathsf{Q}^n$ causally separated, in symbols $p \sim_c q$, iff $\mathsf{time}(p,q) \geq \mathsf{space}(p,q)$, i.e. iff $(p_1 - q_1)^2 \geq (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2$. Visually, $p \sim_c q$ means that q is inside or on the light-cone emanating from p. In $\mathbf{Specrel}_0$ we have $\langle \mathsf{Events}, \sim_c \rangle \cong \langle \mathsf{Q}^n, \sim_c \rangle$, in fact loc_m and ev_m are isomorphisms between these structures, for any $m \in \mathsf{Ob}$.

Let $\mathfrak{M} \models \mathbf{Specrel}_0$ and define its *causal-geometry* as

$$\mathsf{Cg}(\mathfrak{M}) := \langle \mathsf{Events}, \sim_c \rangle.$$

Ants and elephants may use different units of measurement (e.g., their feet). The following axiom expresses that we abstract from the value of the units of measurement. We do so by requiring that all kinds of units of measurement be there. (In connection with the intuition/philosophy related to the following "ant-elephant" axiom cf. the Incredible Shrinking Man in Nicholls, 1982, pp. 194–195.)

$$\mathsf{AxDil} \quad (\forall m \in \mathsf{Ob})(\forall \lambda > 0)(\exists k \in \mathsf{Ob})(\forall p \in \mathsf{Q}^n)\, \mathsf{w}_{mk}(p) = \lambda p.$$

$$\mathbf{Catrel}_0 := \mathbf{Specrel}_0 \cup \mathbf{Comp} \cup \{\mathsf{AxDil}\} = \mathbf{Catrel} - \{\mathsf{AxSim}\} \cup \{\mathsf{AxDil}\}.$$

Assume $\mathfrak{M} \models \mathbf{Catrel}_0$. Then every part of $\mathfrak{M}$ can be recovered from $\mathsf{Cg}(\mathfrak{M})$, except 0 and 1. By this we mean that $\mathsf{B}, \mathsf{Ob}, \mathsf{Ph}, \mathsf{W}, \mathsf{Q}, \leq$ all can be defined (or recovered) from $\mathsf{Cg}(\mathfrak{M})$, but instead of $+, *$, which are not definable, we can define their *affine* ternary versions $+_3, *_3$ where

$$+_3(x,y,z) := x + y - z, \qquad *_3(x,y,z) := x * y/z.$$

Alternately, $\mathfrak{M}$ can be defined over $\mathsf{Cg}(\mathfrak{M})$ parametrically only, i.e. if we add two (arbitrary) constants to $\mathsf{Cg}(\mathfrak{M})$. So we have here an analogue of Theorem 11.16 working between $\mathbf{Catrel}_0$ and its causal-geometries of form $\mathsf{Cg}(\mathfrak{M})$. In particular, $\mathfrak{M}$ and $\mathsf{Cg}(\mathfrak{M})$ are definitionally equivalent in the parametric sense of definability.

Instead of stating the precise analogue of Theorem 11.16 for this intimate connection between causal-geometries and $\mathbf{Catrel}_0$, we turn to the question of what the definable relations in causal-geometries are.

Algebraic logic is a branch of logic that investigates the structure of the definable concepts in a theory, or in a model of a theory. Let us consider $\mathsf{Cg}(\mathfrak{M})$ for an arbitrary $\mathfrak{M} \models \mathbf{Catrel}_0$. The definable relations in $\mathsf{Cg}(\mathfrak{M})$ are exactly those absolute properties of events which do not involve concrete values of the metric $\boldsymbol{\mu}$. We will call these relations *causal-relations*.

The unary (i.e. one-place) causal-relations are Events and $\emptyset$. What are the binary (i.e. two-place) causal-relations?

We call events e, e' *timelike (lightlike, spacelike) separated*, in symbols $e \sim_t e'$ ($e \sim_\ell e'$, $e \sim_s e'$) iff [$e \neq e'$ and $e \cap e' \cap \mathsf{Ob} \neq \emptyset$ ($e \cap e' \cap \mathsf{Ph} \neq \emptyset$, $e \cap e' \cap (\mathsf{Ob} \cup \mathsf{Ph}) = \emptyset$, respectively)]. On the "coordinate-side",

we call space-time locations p, q *timelike (lightlike, spacelike) separated*, in symbols $p \sim_t q$ ($p \sim_\ell q$, $p \sim_s q$) iff [$p \neq q$ and $\mathsf{time}(p,q) > \mathsf{space}(p,q)$ ($\mathsf{time}(p,q) = \mathsf{space}(p,q)$, $\mathsf{time}(p,q) < \mathsf{space}(p,q)$, respectively)]. These are corresponding properties via the bijections loc_m and ev_m for $m \in \mathsf{Ob}$, as before.

Timelike, lightlike, and spacelike separability of events are all causal-relations, i.e. they can be defined from $\sim_c$. In fact, all these four relations can be defined from each other.

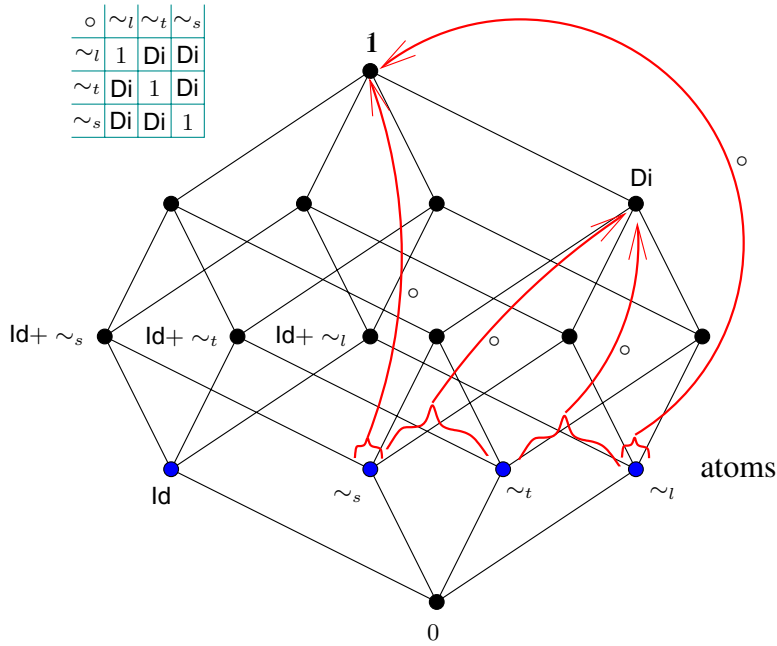
Below we sketch how $\sim_t$ can be defined from $\sim_c$. The argument is easier to follow in the isomorphic structure $\langle \mathsf{Q}^n, \sim_c \rangle$. The points causally separated from a point x are in the "solid" light-cone emanating from x. This light-cone consists of two separate parts, the "upward" and the "downward" parts. We cannot distinguish with a formula the two separate parts of the light-cone, but we can express that "y, z are in the same half-cone of x", in symbols $y \equiv_x z$, as follows.

$$y \equiv_x z \Leftrightarrow [x \sim_c y \wedge x \sim_c z \wedge \exists w(x \sim_c w \wedge \neg w \sim_c y \wedge \neg w \sim_c z)].$$

From this then we can define timelike separability as follows:

$$x \sim_t y \Leftrightarrow \exists zw(x \sim_c z \sim_c y \wedge \neg x \equiv_z y \wedge x \sim_c w \sim_c y \wedge \neg x \equiv_w y \wedge \neg z \sim_c w).$$

We used four variables in the above definition. By using 3 variables only, $\sim_t$ is not definable from $\sim_c$. This can be proved by using the techniques of algebraic logic, as follows. The four relations $\sim_t, \sim_\ell, \sim_s, \mathsf{Id}$ form the atoms of the Boolean algebra they generate. All these relations are symmetric, i.e. they are their own converses. Moreover, the relational composition of any distinct two is $\mathsf{Di} := -\mathsf{Id}$, while the relational composition of any non-identity one with itself is the unit $\mathsf{Events} \times \mathsf{Events}$ of the Boolean algebra. Hence they form a relation algebra. Relation algebras are introduced and briefly discussed in Ch. 3 (see also Henkin et al., 1985, Hirsch and Hodkinson, 2002, Andréka et al., 2001). The elements of a concrete *relation algebra* are binary relations, and the operations are the Boolean ones together with relational composition of binary relations, taking converse of a binary relation, and the relation Id as a constant. The binary causal-relations then form a relation algebra. We can check that in this relation algebra $\sim_t$ is not in the subalgebra generated by $\sim_c$, hence $\sim_t$ cannot be defined from $\sim_c$ by using only three variables, by Tarski's theorem, Tarski and Givant, 1987 or Ch. 3, Proposition 2.4. That $\sim_t$ can be generated from $\sim_c$ by using 4 variables is equivalent to the fact that, in the so-called 4-dimensional *cylindric algebra* of 4-placed causal-relations, $\sim_t$ is indeed in the subalgebra generated by $\sim_c$. For cylindric algebras we refer to Henkin et al., 1985, Henkin et al., 1981, Andréka et al., 2001.



$\circ$	$\sim_l$	$\sim_t$	$\sim_s$
$\sim_l$	1	Di	Di
$\sim_t$	Di	1	Di
$\sim_s$	Di	Di	1

Figure 11.20. The binary causal-relations (in **Specrel**$_0$) form a relation algebra.

One can prove that all the binary causal-relations are the ones occurring in the above relation algebra, which is represented in Fig. 11.20. Thus we have described all the binary causal-relations. Similar definability results for the special case Q =“the rational numbers” are in van Benthem, 1983, pp. 23–30.

There are infinitely many ternary (i.e. 3-place) causal-relations. The most often used ternary and 4-place causal relations are the following ones (again, it is easier to define their “coordinate-versions” in $\langle \mathsf{Q}^n, \sim_c \rangle$).

Collinearity of 3 space-time-locations is a causal relation, let $\mathsf{coll}(p, q, r)$ denote that p, q, r are collinear, i.e. they lie on a straight line.

Betweenness: $\mathsf{Bw}(p, q, r)$ iff “$\mathsf{coll}(p, q, r)$ and q is between p and r”.

Equidistance: $\mathsf{Eq}(p, q, r, s)$ iff “$\boldsymbol{\mu}(p, q) = \boldsymbol{\mu}(r, s)$”. Minkowski-circles (or Minkowski-spheres) can be defined from equidistance, cf. Fig. 11.21.

Orthogonality: $\mathsf{Ort}(p, q, r, s)$ iff “the line connecting p, q is Minkowski-orthogonal to the line connecting r, s”, i.e. iff $(p_1 - q_1)(r_1 - s_1) - (p_2 - q_2)(r_2 - s_2) - \cdots - (p_n - q_n)(r_n - s_n) = 0$. For illustration see Fig. 11.22.

Betweenness with ratio ρ: Let ρ be any rational number. Then $\mathsf{Bw}_\rho(p, q, r)$ iff “$\mathsf{Bw}(p, q, r)$ and $\boldsymbol{\mu}(p, q) = \rho * \boldsymbol{\mu}(q, r)$”.

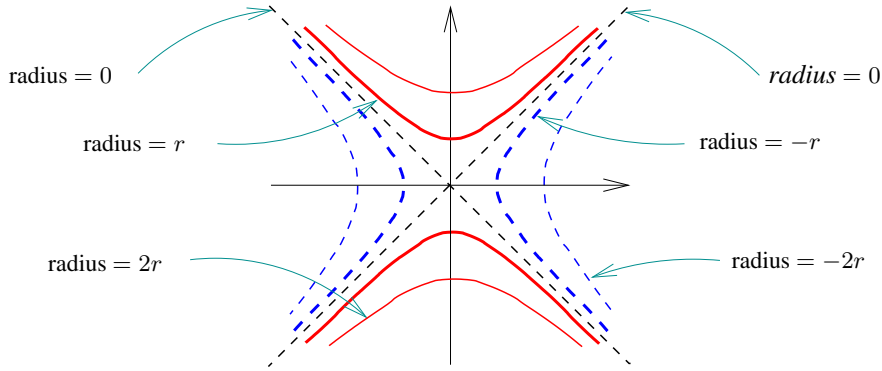


Figure 11.21. Minkowski circles.

The above are all definable from $\sim_c$. The last example shows that there are infinitely many ternary causal-relations.

If we have time-orientation, i.e. in models of $\mathbf{Specrel}_0 + \mathsf{Ax}\uparrow$, the following important relation can also be defined. Event e *causally precedes* event e', in symbols $e \prec_c e'$ iff ($e \sim_c e'$ and $(\exists m \in \mathsf{Ob})\mathsf{time}_m(e) \leq \mathsf{time}_m(e')$). In $\mathbf{Specrel}_0 + \mathsf{Ax}\uparrow$, $e \prec_c e'$ is equivalent with the simpler formula $(\forall m \in \mathsf{Ob})\mathsf{time}_m(e) \leq \mathsf{time}_m(e')$, assuming $n > 2$. It can be proved (analogously to Corollary 11.5) that the corresponding property in space-time locations is: $p \prec_c q$ iff ($p \sim_c q$ and $p_t \leq q_t$). $\prec_c$ is also called *causality relation*, or *after*, the first axiomatization of special relativity in Robb, 1914 axiomatized this causality relation. The general relativistic version of $\prec_c$ is quite important, too, and is more intricate than the $\mathbf{Specrel}_0$ version; cf., e.g., works of Penrose, Malament, Busemann.

Finally, we list some absolute relations that can be defined in $\mathsf{Mg}(\mathfrak{M})$, but cannot be defined in $\mathsf{Cg}(\mathfrak{M})$, for $\mathfrak{M} \in \mathbf{Catrel}$. Clearly, $\boldsymbol{\mu}$ is such.

Minkowski scalar-product: $\mathsf{g}_4(p,q,r,s) := (p_1 - q_1) * (r_1 - s_1) - (p_2 - q_2) * (r_2 - s_2) - \cdots - (p_n - q_n) * (r_n - s_n)$. We note that $\mathsf{g}_4(p,q,p,q) = \boldsymbol{\mu}(p,q)$ and $\mathsf{g}_4(p,q,r,s) = 0$ iff $\mathsf{Ort}(p,q,r,s)$. Hence, g_4 "codes" both Minkowski-distance and "relativistic angle".

Relativistic (non-squared) distance of causally separated points: $\mathsf{rd}(p,q) := \sqrt{(p_1 - q_1)^2 - (p_2 - q_2)^2 - \cdots - (p_n - q_n)^2}$. This is a partial function defined exactly when the expression in the argument of the square root is nonnegative, i.e. when $p \sim_c q$. $\mathsf{rd}(e,e') > 0$ means proper time elapsed between e and e', for any observer who takes part in both e and e' (and that there is such an observer), i.e. any observer who takes part in both e, e' measures that the elapsed time between e and e' as $\mathsf{rd}(e,e')$ (and there is such an observer).

We note that $\boldsymbol{\mu}, \mathsf{g}_4, \mathsf{rd}$ are definable from each other. Actually, we will make use of this in our section on general relativity, in Sec. 3.4, where we will use the binary version g of g_4 which is defined by

$$\mathsf{g}(p,q) := \mathsf{g}_4(p,\overline{0},q,\overline{0}).$$

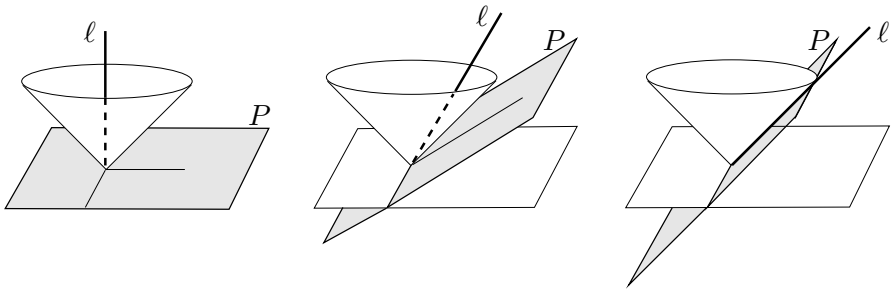


Figure 11.22. ℓ is Minkowski-orthogonal to each straight line in P.

For more concrete definitions and for intuition for the above relations we refer to e.g. Goldblatt, 1987, or Madarász, 2002, Sec. 4.2.

By Theorem 11.16 and the analogue for **Catrel**$_0$, if we want to study special relativity in the form of **Catrel** or **Catrel**$_0$, then this can be done *equivalently* by studying the simple metric-geometries and causal-geometries $\langle \mathsf{Events}, \boldsymbol{\mu}; \mathsf{Q}, 1\rangle$ and $\langle \mathsf{Events}, \sim_c\rangle$, using their finitely axiomatized theories Th_m and Th_c respectively. (A finite Th_c can be found in Latzer, 1972.) The possibility of switching to the geometries instead of the original models (without losing information) is useful e.g. because the transition from special relativity to general relativity (GR) is quite smooth on the level of geometries. The study of GR on this causal-geometric (axiomatic) level of abstraction is promoted e.g. in Kronheimer and Penrose, 1967. Busemann, 1967 uses the generalization of rd in his approach to general relativity space-times. The study of GR on the metric-geometric level using the generalization of g is most common, see e.g. Wald, 1984, Hawking and Ellis, 1973, Rindler, 2001 (but in this "g-oriented" cases the linguistic economy of the Kronheimer-Penrose approach is usually sacrificed).

One of the most useful and most interesting branches of mathematical logic is, in our opinion, definability theory. Definability theory is strongly related to relativity theory, in fact its existence was initiated by Hans Reichenbach in 1924 (Reichenbach, 1969) motivated by relativity theory. Reichenbach in his works emphasized the need of definability theory and made the first steps in creating it. It was Alfred Tarski who later (1930) founded and established this branch of mathematical logic.

Very briefly, the reason for the need of definability theory (of logic) in relativity theory is as follows. When one sets up a physical theory Th, one wants to use only so-called observational concepts, like e.g., "meeting of two particles".[7] While investigating the theory Th, one defines new, so-called "theoretical" concepts, like e.g. "relativistic distance of events". Some defined concepts then prove to be so useful that one builds a new theory Th' based on the most useful theoretical concepts, and investigates this new theory Th' in its own merits.

The new theory Th' usually is simple, streamlined, elegant—built so that we satisfy our aesthetic desires. The original theory Th contains its own interpretation, because we defined it so. The physical interpretation of the new streamlined theory Th' is provided by its connection with Th. The strongest known relationship between two theories is definitional equivalence. When Th and Th' are definitionally equivalent, in the rigorous sense of definability theory of first-order logic, the observational oriented theory Th can be recaptured completely from the theoretical-oriented streamlined theory Th'; and vice versa, the theoretical concepts of Th' can be defined (justified) over the observational Th. Looser relationships between Th and Th' are also very useful, these kinds of relationships between theories are called interpretability and duality theories. Cf. van Benthem, 1982 for more on logic, definability theory, model theory for empirical theories.

In Secs. 2.1–2.3 when we formalized special relativity in first-order logic, we tried to choose the basic concepts of our language as observational as possible; and we introduced the more theoretical concepts of relativity as definitions at later stages, when development of the theory justified them. Eventually, in Sec. 2.6, this process lead to the introduction of a new theory with new basic concepts (new vocabulary, like $\sim_c$, $\prec_c$). This is a natural way of theory development, theory "understanding", theory analysis. In modern approaches to logic, theories are considered as *dynamic objects* as opposed to the more classical "eternally frozen" idea of theories. For approaches to the dynamic trend in mathematical logic cf. van Benthem, 1996.

[7] The concepts potentially usable in scientific theories (such as e.g. relativity) have been partially ordered in the literature as being more observable (and less "theoretical") or less observable and more theoretical. Here "observable" also means primary or empirical. This observable/theoretical distinction, or rather hierarchy, is recalled from the literature (of relativity theory) in e.g. Friedman, 1983, pp. 4–5. This observable/theoretical hierarchy is not perfectly well defined and is known to be problematic, but as Friedman puts it, it is still better than nothing. E.g. the motion of the oceans called tides are more observable (or closer to be observable) than the pull of gravity of the Moon which, we think, is causing them. That is, the gravitational force field of a mass-point (like the moon) is a more theoretical concept than the motion of a body (e.g. ocean's shore-line). Actually the gravitational force field might turn out to be a "wrong concept" and we may have to replace it with something else like the curvature of space-time. Probably the motion of the ocean's shore-line will be less questionable as a "something" which one can talk about. As Friedman, 1983, p. 4 points out, the observational/theoretical distinction is not an absolute one. E.g. what is an observational concept at a certain stage of theory development might turn out to be a theoretical one later.

Theories form a rich structure when we investigate their interconnections. Algebraic logic establishes a duality between hierarchies of theories (on different vocabularies) and between classes of algebras, cf. e.g., Henkin et al., 1985, Sec. 4.3 or Andréka et al., 2001, Part II. Investigating a theory via investigating the hierarchy of its different perspectives and subtheories is like investigating a 3-dimensional object from all sides. This leads us to the subject of the next subsection.

2.7 Conceptual analysis and "reverse relativity"

In the previous subsections we strengthened **Specrel** to a complete theory **Catrel**+*RC*. But in reality, when working with a theory, we do not want to make our axioms generate a complete theory. Our purpose is just the opposite: we want to make our axioms as weak, simple, and intuitively acceptable and convincing as possible while still strong enough for proving interesting theorems of relativity theory. Similar striving for economy of assumptions is e.g. in Szabó, 2006, Szabó, 2002, Ax, 1978. The reasons for wanting to study weak theories as opposed to strong ones are, among others, the desire for answering the "why-type questions", and seeking a conceptual analysis of the theory. For more on this we refer to Andréka et al., 2002, Sec. 1.1. Further reasons for striving for weak physical theories having many models are presented in van Benthem, 1982. Namely, e.g. "small" mechanical systems like our solar system or another one or our galaxy can be regarded as many different "small models" of e.g. Newtonian mechanics.

Among other things, we can use logic to find out which axioms are responsible for certain surprising predictions of relativity theory like e.g. "no observer can move faster than the speed of light", "the twin paradox" or issues concerning the possibility of time travel. We can call such studies "reverse relativity" alluding to the analogy with the highly successful direction called reverse mathematics, cf. e.g. Simpson, 2005, Friedman, 2004.

In reverse relativity, we single out an interesting prediction of relativity theory like "observers cannot move faster than light (NoFTL)", or one of our paradigmatic effects in Theorem 11.6 and ask ourselves which axioms (of e.g. **Specrel**) are responsible for the prediction in question, cf. Theorem 11.7. Let φ denote the prediction in question. So typically we know that **Specrel** $\vdash \varphi$ and that φ "is interesting". Then we ask ourselves whether the whole of **Specrel** is needed for proving φ. (Recall, the axioms of **Specrel** are "assumptions", hence they cost money so to speak.) A further question is to ask which fragment of **Specrel** is needed/sufficient for proving φ. This type of research has been carried through for several interesting choices of φ, e.g. in Andréka et al., 2002, Madarász, 2002, Andréka et al., 2004.

Let us take as an example for φ the prediction NoFTL (i.e. that no observer can move faster-than-light relative to another) established as Theorem 11.7(ii), p. 636. Certainly, NoFTL is an interesting prediction, indeed, many thinkers tried to get rid of NoFTL either by using "tachions" or by circumnavigating it by using wormholes, cf. Sec. 4 for the latter. An instructive "saga" of such efforts is provided in Thorne, 1994. The analysis of NoFTL tells us that **Specrel** can be considerably weakened without losing NoFTL. By Theorem 11.7 in this chapter, the assumption $n > 2$ is needed, however. The two key axioms of **Specrel** are the Light Axiom, AxPh, and AxEvent (in some sense). It turns out that both of these are needed for NoFTL. However, both of them can be weakened considerably without losing NoFTL. In case of AxPh, isotropy is not needed for NoFTL. Of AxPh, it is enough to assume that photons are not like bullets, they do not race with each other, and they can be sent from each point in each direction; i.e. that for any observer m in each direction d there is a number $c_m(d) \in \mathsf{Q}$ representing the speed of light for m in direction d. Of AxEvent, it is enough to assume that if m sees an event on the worldline of k, then k also sees that event; and that if m sees an event that k sees then m sees all events in a neighborhood (in k's coordinate system) of this event. Some reflection reveals that this is a more natural, milder assumption than AxEvent was. As it turns out, the rest of the axioms of $\mathbf{Specrel}_0$ can also be weakened without losing NoFTL.

Careful analysis of the noFTL prediction can be found in Madarász, 2002, 2.8.25, 3.2.13, 3.2.14, Madarász et al., 2004, Theorem 3, Theorem 5, Andréka et al., 2002. Similar pieces of conceptual analysis, analysing predictions similarly interesting (like NoFTL) can be found in Andréka et al., 2002, Sec. 4.2, Andréka et al., 2004, Madarász et al., 2004, Madarász et al., 2006b, Madarász et al., 2006a. Predictions that have been analyzed in these works include the twin paradox, the paradigmatic effects, the effect of gravity on clocks.

3. General relativistic space-time

In this section we extend our logic-based study of relativity from special relativity to general relativistic space-time (GR space-time). In particular, in Sec. 3.6 we present a purely first-order logic axiomatization **Genrel** for GR space-time. Theorem 11.28 is a kind of completeness theorem for **Genrel**. Besides providing a first-order logic axiomatization of GR space-times (analogously to Sec. 2) and comparing it with **Specrel**, we will put extra emphasis on discussing the exotic properties of various distinguished examples of GR space-times in Sec. 4. One of the reasons for this is that these exotic GR space-times are at the center of attention nowadays, e.g. because of their fantastic properties and because astronomers have been discovering examples of these, e.g. finding observational evidence for huge black holes in the last 15 years.

Another reason for putting emphasis on examples is that while **Specrel** has basically one intended model, general relativity (GR) has many different intended models (e.g. various kinds of exotic black holes, wormholes, timewarps, models for the expanding universe, the Big Bang, to mention a few). This contrast between special relativity and GR motivates our shifting the emphasis to distinguished models in what comes below. This kind of motivation is further elaborated in Taylor and Wheeler, 2000.

A motivation for the logical analysis of GR is that, in principle, GR space-times permit such counter-common-sense arrangements as is time travel (in one form or another). This was discovered by Kurt Gödel during his cooperation with Einstein. But the so-called paradoxes of time travel offer themselves for a logical analysis, since these kinds of circularity are the "bread-and-butter" of the logician ever since Gödel's incompleteness proof or since the first logical analysis of the liar paradox and its variants. Even if we would want to exclude time travel by some axiom like one or another form of the so-called Cosmic Censor Hypothesis, it remains a question how to find and justify a natural axiom to this effect without making unjustified assumptions. This dilemma is illustrated by the debates about the various forms of the Cosmic Censor Hypothesis and related assumptions discussed e.g. in Earman, 1995.

3.1 Transition to general relativity: accelerated observers in special relativity

In **Specrel**, we restricted attention to inertial observers. It is a natural idea to generalize the theory to including accelerated observers as well. Actually, when creating general relativity, Einstein emphasized that accelerated observers should be included, cf. Einstein, 1961, pp. 59–62. Indeed, the usual transition from special relativity to the general theory of relativity goes as follows. First special relativity is generalized to accommodate accelerated observers, and then one introduces *Einstein's principle of equivalence* (EPE) which states that the phenomena of acceleration and gravity are equivalent (in a carefully specified concrete sense). Then, at this point, our language is rich enough to talk about gravity in the form of acceleration. After this point, one refines the theory, arriving at GR, and then it all hangs together to form a worldview broader than special relativity and also broader than Newtonian gravitation theory. The above is illustrated by e.g. Einstein, 1961, the classic general relativity book Misner et al., 1970, pp. 163–165, Rindler, 2001, e.g. p. 72, §3.8, §12.4, pp. 267–272. Even works intending to venture to the unknown beyond GR use the above "methodology" of starting by accelerated observers, cf. e.g. Smolin, 2001, pp. 77–80.

The same is done in the research area reported in the present work. Namely, in Sec. 2 and in related works, the logical analysis of special relativity is done, yielding e.g. the hierarchy of theories containing **Specrel**$_0$, **Specrel**, **Catrel**. The next stage extends **Specrel** by considering new kinds of entities called accelerated observers and stating further axioms governing their behavior. This yields a new theory **Accrel** which can be regarded as an extension and refinement of **Specrel**. Gravity can be studied in **Accrel** in form of acceleration; this is done e.g. in Madarász et al., 2006a, in the spirit outlined above. The works Andréka et al., 2006b, Madarász et al., 2006b, Madarász et al., 2006a which study **Accrel** stay inside the purely first-order logic based approach represented by **Specrel** in Sec. 2 of this chapter. Using the experience and motivation gained by studying **Accrel**, in Sec. 3.6 we introduce a first-order logic theory **Genrel** for general relativistic space-time. All this converges to a logical analysis of GR.

Instead of recalling **Accrel**, which is very similar in spirit to the axiom system **Genrel** in Sec. 3.6, we summarize its main features relevant to **Genrel**. In **Accrel**, "accelerated observer" means "not necessarily inertial observer", and "observer" means a body that has a worldview, i.e. which occurs in the domain of the worldview relation W. Thus, an accelerated observer has a worldview. Roughly, the worldview of an accelerated observer k is obtained from the worldview of an inertial one, m, by re-coordinatizing it along a smooth bijection w_{mk} with an open subset of Q^n as its domain. Thus k may use only part of Q^n for coordinatizing events, and more importantly, the worldlines of inertial observers and photons are no longer straight lines in an accelerated observer's worldview. I.e., $\mathsf{AxEvent}$ and AxLine cease to hold. They hold in generalized, weaker forms only. Specifically, an accelerated observer can recognize worldlines of inertial bodies as so-called "geodesic curves", this is the motivation for Sec. 3.3 and for the axiom AxLine^- in **Genrel**.

The key axiom of accelerated observers states that at each moment of his life, each accelerated observer sees the nearby world for a short while as an inertial observer does. Technically, in **Accrel** we formulate this as stating that at each moment of the life of an accelerated observer k there is a so-called co-moving inertial observer m such that the linear approximation (i.e. the differential) of the worldview transformation w_{mk} at this space-time point is the identity function. This axiom, called AxAcc in the quoted works, is the connecting point between the worldviews of inertial and non-inertial observers. The counterpart of AxAcc in the present work will be discussed next.

We can think of an accelerated observer in special relativity as a spaceship which uses fuel for accelerating (in a space where there is no gravity). When the drive is switched off, the ship will transform into an inertial ship—this is the co-moving inertial observer at the event of switching off the ship-drive. Or, equivalently, we can think of the co-moving inertial observer as a spaceprobe which was let go from the ship—a metaphorical apple dropped in space. The

ship can measure its own acceleration by measuring the acceleration of the spaceprobe; just as here on Earth we can measure gravity by measuring the acceleration of a dropped apple. EPE then implies that the spaceship can interpret "falling of the metaphorical apples" either by thinking that he is accelerating in an empty space, or by thinking that he is suspended in a space where there is gravity, and dropped apples fall because they are no longer suspended. By EPE, the worldview of an accelerated observer in special relativity is similar to a "suspended" observer in a space-time where there is gravity. The gravitational counterpart, by EPE, of AxAcc is Einstein's Locally Special Relativity Principle which we recall at the beginning of Sec. 3.2; it will be our starting point in defining GR space-times.

Summing up: by EPE, investigation of gravity can be reduced to the investigation of the worldlines of inertial bodies in a GR space-time.

Let us turn to the reasons of why the transition from special relativity to general relativity goes via accelerated observers and EPE. In Sec. 2, we chose to derive special relativity from the outcome of the famous Michelson-Morley experiment, i.e. from the Light Axiom. However, as we already mentioned, relying on the Light Axiom is not really necessary. As Einstein always emphasized (e.g. in Einstein, 1961), relativity can be derived from a deep philosophical principle called the *special principle of relativity* (SPR). SPR has been around in our culture for 2500 years (roughly), hence it is well understood and it blends nicely with our best understanding of the world. Roughly, SPR says that the Laws of Nature are the same for all inertial observers. The modern form of SPR was articulated by the Normann-French Nicole d'Oresme around 1300 (Paris) and (a bit more thoroughly) by Galileo Galilei (around 1600). After Olaf Roemer, James Bradley and followers discovered that the speed of light is finite (and related issues were clarified), SPR could have been used[8] to show inconsistency with the Newtonian worldview and then to derive special relativity (analogously to the train of thought we used in Sec. 2). This in turn would have predicted the outcome of the Michelson-Morley experiment.[9] Einstein elaborated this idea in detail, and in particular, emphasized that special relativity can be derived from SPR (in place of the Light Axiom). The same kind of philosophical taste led Einstein to asking why SPR is restricted to inertial observers only. Why aren't the Laws of Nature the same for all observers (not only the inertial ones)? After all, we ourselves sitting on the surface of the Earth (and fighting gravity all the time) are not inertial observers according to the definition used in SPR.

[8]With hindsight, this possibility was there around the 1830's or so. Roemer made the discovery around 1680, but it was not generally accepted until 1750 approx.

[9]Of course, for this, light propagation needs to be regarded as Law of Nature, but as Einsteinpoints out, this is absolutely natural.

So, Einstein started working towards GR by generalizing SPR to the *general principle of relativity* (PR) which says that the Laws of Nature are the same for all observers, including accelerated ones. This move creates some extra tasks to handle, because accelerated observers experience the existence of a new "force-field", namely gravity. So Einstein introduced his principle of equivalence EPE unifying acceleration created by gravity with "ordinary gravity". Now, to uphold PR we have to regard gravity (and light propagation of course) as part of what constitute Laws of Nature. This creates some extra tasks (mentioned above) since in **Specrel** properties of gravity were not part of the picture. As we will see in the next section, this extra work can be handled leading to a unification broader than that provided by special relativity. The new theory GR unifies space, time, motion, light-propagation, and gravitation into a single purely geometrical perspective.

3.2 Einstein's "locally special relativity principle"

Einstein's locally special relativity principle saying that General Relativity is locally Special Relativity is the following. Let p be a point in a GR space-time. Then if we drop a small enough spaceship, put an experimental scientist in the spaceship who lives for a short enough time, the experimentalist will find special relativity true in the spaceship. This holds true even on the event horizon of a spinning black hole or wherever you want. Of course, it is crucial that the spaceship is small enough and that its life is considered only for a small enough time interval. This is Einstein's locally special relativity principle. See Fig. 11.23. These local tiny spaceships will appear later as "local reference frames" LFR's. They play the same role in GR as co-moving observers did in AxAcc.

Next we implement Einstein's locally special relativity principle formulated above for formalizing GR space-times. In this and in the next few subsections, for simplicity, we will use $\mathbb{R}$ in place of an arbitrary linearly ordered quadratic field Q, and also we will use $n = 4$. Later, in Sec. 3.6 we will return to the generality of Q and $n \geq 2$.

For general relativity, we will use *global coordinate frames, GFR's*. A global coordinate frame is based on an open subset of $\mathbb{R}^4$. So a global coordinate frame looks like a special relativity frame, we even call one of the coordinates time, the others space etc. The difference is that in a global frame the coordinates do not carry any physical or intuitive meaning.[10] They serve only as a matter of convention in gluing the local special relativity frames, LFR's, together. For simplicity, at the beginning we will pretend that the coordinate system of our global frame is the whole of $\mathbb{R}^4$. Later we will refine this to saying that the

[10]To be precise, the topology of the global frame will be relevant.

global frame is an open subset of $\mathbb{R}^4$. And even later, in Sec. 3.6, we will generalize this to be a manifold. Since the differences are extremely minor and secondary from the point of view of the basic notions we are going to introduce now, let us first pretend that the global frame is $\mathbb{R}^4$.

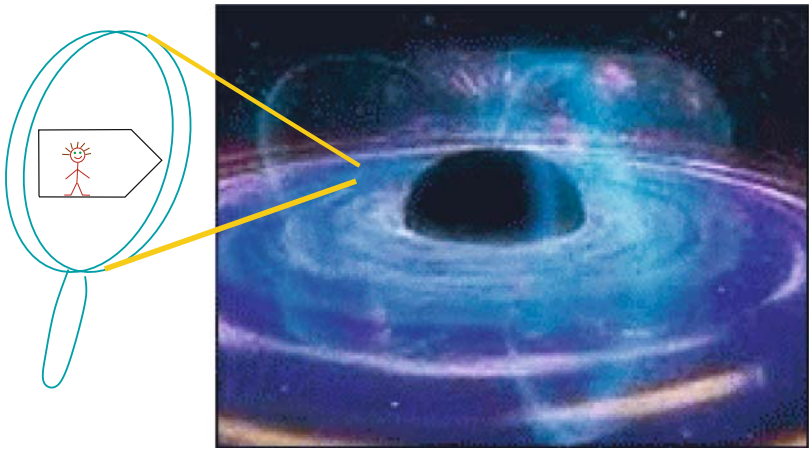


Figure 11.23. Einstein's locally special relativity principle: where-ever we drop a small enough spaceship, for a short enough time it will experience special relativity.

Imagine a general relativistic coordinate system, a GFR, representing the whole universe, with a black hole in the middle etc. So we are looking at the bare coordinate grid of $\mathbb{R}^4$ intending to represent the whole of space-time. What is the first thing we want to specify for our readers about the points of this grid $\mathbb{R}^4$? Well, it is how the local tiny little special relativistic space-times are associated to the points p of $\mathbb{R}^4$, in accordance with Einstein's locally special relativity principle formulated at the beginning of this subsection. Thus, to every point p of $\mathbb{R}^4$ we want to specify how the local special relativity space-time at point p is squeezed into the local neighborhood of p. The point is in specifying how the clocks of the LFR slow down or speed up at p, and which axis of the local LFR points in what direction and is distorted (shortened/lengthened) in what degree. The LFR at p corresponds to the metaphorical spaceship dropped at p as in Fig. 11.23.

How do we specify the local frames LFR's? A *local frame* L_p at p will be a bijective mapping $L_p : \mathbb{R}^4 \to \mathbb{R}^4$ such that $L_p(\overline{0}) = p$. We will think of the first $\mathbb{R}^4$ as the coordinate system of special relativity, or of the Minkowski space represented by LFR, and of the second $\mathbb{R}^4$ as the global frame GFR upon which we want to build our GR space-time. Thus we write $L_p : \text{LFR} \to \text{GFR}$, where formally $\text{LFR} := \text{GFR} := \mathbb{R}^4$. Since we want to use our local frames to specify how the tiny clocks slow down in the "linear limit" (roughly, in an

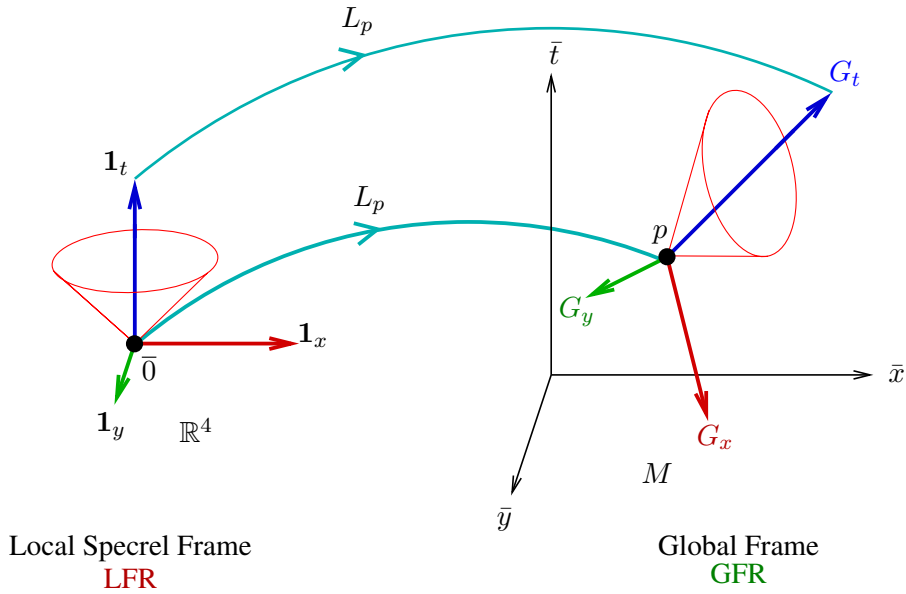


Figure 11.24. The local frame at p is an affine mapping L_p of $\mathbb{R}^4$ to $\mathbb{R}^4$ taking $\bar{0}$ to p. We will use L_p in small neighborhoods of $\bar{0}$ only.

"infinitesimally small" neighborhood of p), we will choose these L_p's to be affine mappings.

So, the key device in building our GR space-time is associating to each point p of our global coordinate grid GFR an affine transformation L_p mapping the Minkowski space represented by LFR to the global frame GFR.

DEFINITION 11.18 (GENERAL RELATIVISTIC SPACE-TIME) *By a* general relativistic space-time *we understand a pair* $\langle M, L\rangle$ *where*

$M \subseteq \mathbb{R}^4$ *is open and*

L *is a function defined on* M *such that* $L(p) : \mathbb{R}^4 \to \mathbb{R}^4$ *is a bijective affine mapping which takes the origin* $\bar{0}$ *to* p*, for each* $p \in M$*. Further, we require that* L *(as a function of* $p \in \mathbb{R}^4$*) be infinitely many times continuously differentiable, in short* smooth.

For better readability, we will write L_p to denote $L(p)$. We will use local special relativity frames (the LFR's) for importing the notions of special relativity to our GR space-times. We will use the inverse mapping L_p^{-1} of the affine transformation L_p to translate our general relativistic problems to special relativity, and we will use L_p to bring back the answers special relativity gives us. Though L_p^{-1} is defined on the whole of $\mathbb{R}^4$, we will use it only in small enough

neighborhoods of p (i.e. we will use it in the limit, more and more accurately as we close on p). Restricting attention to small neighborhoods of p is what is meant by saying that GR can be *locally* reduced to special relativity, but only locally. If we want to solve a problem at a point q farther away from p , then we will have to use the mapping L_q associated to q in place of using L_p.

We can specify the local frame L_p by the images of the four unit-vectors $\mathbf{1}_i$, as follows: $G(p) = \langle G_t(p), \ldots, G_z(p)\rangle$ is a 4-tuple of vectors (i.e. elements of $\mathbb{R}^4$) such that $L_p : \mathbb{R}^4 \to \mathbb{R}^4$ is the affine transformation which maps the origin $\overline{0}$ to p, and $\mathbf{1}_i$ to $G_i(p) + p$, for $i \in \{t, x, y, z\}$, see Fig. 11.24. Thus, we can specify a GR space-time $\langle M, L\rangle$ by simply specifying four vector-fields. Here we use the word "field" as in analysis and not as in algebra. Thus "field" in "vector-field" means that we have a vector at each point of $M \subseteq \mathbb{R}^4$. This gives us the following equivalent definition for a GR space-time.

DEFINITION 11.19 (GENERAL RELATIVISTIC SPACE-TIME IN VECTOR-FIELDS FORM) *By a* general relativistic space-time in vector-fields form *we understand a four-tuple* $\langle G_t, G_x, G_y, G_z\rangle$ *of vector-fields such that*

each $G_i : M \to \mathbb{R}^4$ *is smooth, where* $M = \mathsf{Dom}(G_i) \subseteq \mathbb{R}^4$ *is open* ($i \in \{t, x, y, z\}$) *and*

the vectors $G_t(p), \ldots, G_z(p)$ *at each point* $p \in M$ *are linearly independent in the usual sense. (This means that the affine mapping they specify is a bijection.)*

An advantage of Def. 11.19 is that it is simple, and that it is in this form that we can "draw" or visualize a general relativistic space-time. See Figs. 11.30, 11.31, 11.33, 11.40, 11.34.

Now, how do we use a GR space-time, i.e. such a 4-tuple of vector-fields, for representing some aspects of reality? Very, very roughly, the information content of a GR space-time $G = \langle G_t, \ldots, G_z\rangle$ can be visualized as follows. The vector tetrad $G_t(p), \ldots, G_z(p)$ at point p tells us how the measuring instruments (clocks, meter-rods) of the tiny little inertial observer we imagine as being dropped at p go crazy (go wrong) from the point of view of the big, global general relativistic coordinate grid GFR we are using in G. This information is very subjective, since as we said, the big global coordinate grid carries no physical meaning, it is conventional. But some objective content can be extracted from this subjective information. As we said, for a fixed p, the vector tetrad $G_t(p), \ldots, G_z(p)$ wants to represent how the local frame is glued into the holistic picture of the global frame. The important point is, however, how the individual local frames are distorted, rotated etc w.r.t. *each other*, the big global frame grid is only a theoretical, conventional device to serve as a common denominator in arranging the little local frames relative to each other.

What can be described in terms of the 4-tuple $\langle G_t, \ldots, G_z \rangle$ of vector-fields? Well, we can describe the (potential) worldlines (parameterized with wristwatch times) of inertial observers and the worldlines of photons. We will see that knowing what the potential worldlines of inertial observers and photons are tells us everything important about a GR space-time. Gravity, curvature can be defined explicitly from knowing the above mentioned worldlines.

3.3 Worldlines of inertial observers and photons in a general relativistic space-time

The worldlines of inertial observers will be described mathematically as timelike geodesic curves. We now turn to defining these. In this subsection we fix a general relativistic space-time $\langle M, L \rangle$, and we let GFR $:= M$, LFR $:= \mathbb{R}^4$.

By a (smooth) *curve* f we understand a smooth mapping $f : I \to \mathbb{R}^4$, where I is an open interval of $\mathbb{R}$. By a point of the curve we mean a point in its range.

Intuitively, the curve $f : I \to$ GFR is called *timelike* at point p iff the local frame at p "sees" an observer co-moving with the curve at p. In more detail, the curve f is called timelike at p iff the speed of f as seen by the local frame at p is smaller than 1. This means that the tangent of the curve $L_p^{-1} \circ f$ has slope smaller than 1, at the origin; geometrically this means that the tangent straight line at the origin is within the light-cone. The curve f is called timelike iff the curve f is timelike at each of its points p . (Note: this is independent of how the curve f is parameterized.) See Fig. 11.25.

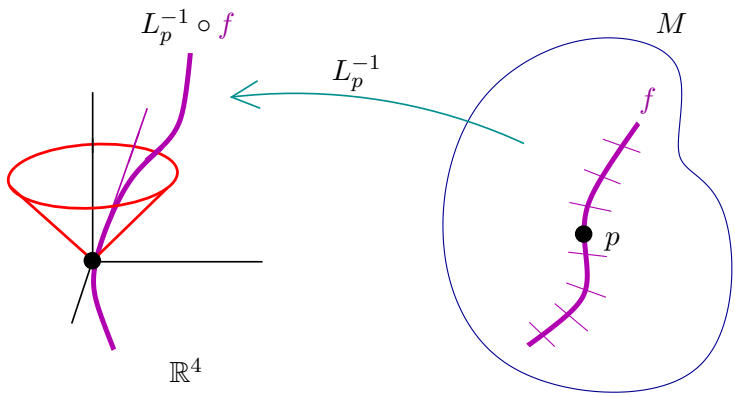


Figure 11.25. The curve f is timelike at point p.

Note that talking about the tangent of $L_p^{-1} \circ f : I \to$ LFR involves nothing "fancy", since (at this step) we are in a special relativity space-time and we are using its Euclidean geometry over $\mathbb{R}^4$ and we are looking at a smooth curve in it.

We think of a timelike curve as a curve that in principle can be the worldline of a (perhaps accelerated) observer.

When is a timelike curve $f : I \to \text{GFR}$ called a timelike geodesic? First we have to check whether the curve f represents (or measures) relativistic time correctly. Here, the parametrization will be important. In what follows, $[a, b]$ denotes the closed interval of $\mathbb{R}$ with endpoints a, b, i.e. $[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}$. For the definition of relativistic distance $\mathsf{rd} : \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}$ see Sec. 2.6, p. 656.

DEFINITION 11.20 *We say that f* represents time correctly *if the following statement holds. For every $t \in I$, and for every positive ε there is a positive δ such that for all $s \in [t - \delta, t + \delta]$ it holds that $|s - t|$ agrees with what $\mathsf{rd}(f(s), f(t))$ is as measured by the local frame determined by $L_{f(t)}$ up to an error bound by $\varepsilon * |s - t|$. Here "$\mathsf{rd}(f(s), f(t))$ as measured by the local frame" is the relativistic (Minkowski) distance between $L_p^{-1}(f(s))$ and $L_p^{-1}(f(t))$ understood in special relativity. Formally this is:*

$$(\forall t \in I)(\forall \varepsilon > 0)(\exists \delta > 0)(\forall s \in [t - \delta, t + \delta])$$
$$|\mathsf{rd}(L_p^{-1}(f(s)), L_p^{-1}(f(t))) - |s - t|| < \varepsilon * |s - t|.$$

We call a curve *time-faithful* iff it is timelike and represents relativistic time correctly. Intuitively, a timelike curve is time-faithful iff at each point p of the curve, the local frame at p "sees" an observer co-moving with the curve such that the parametrization of the curve "agrees" with how time passes for this co-moving observer. See Fig. 11.26.

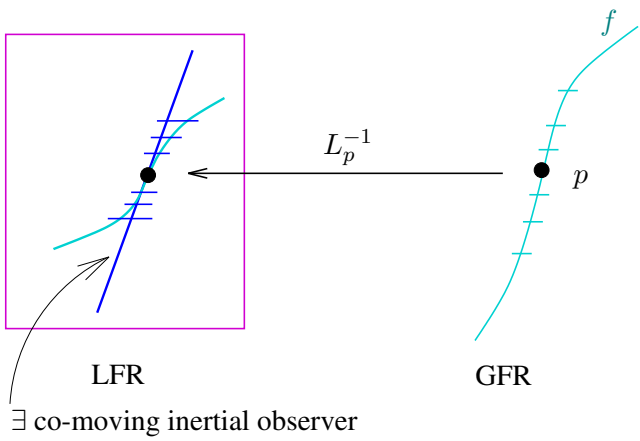


Figure 11.26. A time-faithful curve at p.

We imagine that a time-faithful curve f is the worldline of an observer b such that the parameter t measures proper time of b; or in other words, t shows the time on the wristwatch of b. We imagine the motion of b such that $f(t)$ in GFR is the space-time location of observer b at his wristwatch time t. The condition in Def. 11.20 serves to ensure that wristwatch time t of observer b (whose motion is represented by the curve f) agrees with the relativistic time interval measured by the relativistic metric rd at the local frame which is situated at the location $f(t)$. More precisely, small time intervals on the wristwatch of b agree with the relativistic time interval measured by the rd's of the local special relativity frames. Thus f is the general relativistic analogue of the special relativistic $\mathsf{wline}_m(b)$ + parametrization with "proper time" or "inner time" of b.

Put differently, we think that a timelike curve can be the worldline of (the mass-center of) a spaceship which uses fuel (i.e. uses its ship-drive) for accelerating and decelerating. The curve is time-faithful if the parametrization of the curve agrees with "inner time" of the spaceship.

DEFINITION 11.21 (TIMELIKE GEODESIC) *By a* timelike geodesic *we understand a time-faithful curve $f : I \to$ GFR satisfying the following. For any t in I there is a neighborhood S of $f(t)$ (understood in $\mathbb{R}^4$) such that inside S, f is a "straightest possible" curve in the following sense: For any two points p, q of S connected by f, the distance of p and q as measured by f is maximal among the distances measured by "competing" time-faithful curves inside S.*

Formally, this maximality condition is expressed by the following. Assume that h is a time-faithful curve with range inside S. Assume that $p = f(s) = h(s')$, $q = f(r) = h(r')$ and $h(t')$ is in S for all t' which are between s' and r'. Then $|s - r| \geq |s' - r'|$. See Fig. 11.27.

If f is a timelike geodesic, then we imagine that an inertial observer b can move along it. By an inertial observer b we imagine (the mass-center of) a spaceship with ship-drive switched off, i.e. a spaceship which does not use fuel for influencing its motion.

The above definition of timelike geodesics is the natural reformulation of the Euclidean notion of geodesics (straight curves in a possibly curved surface of Euclidean 3-space) with "minimal" replaced by "maximal" (cf. Hawking and Ellis, 1973, Proposition 4.5.3, p. 105). If one thinks about this maximality condition, one will find that it is strongly connected to the Twin Paradox of special relativity. Indeed, the Twin Paradox is exactly the statement that in special relativity, worldlines of inertial observers are timelike geodesics in the sense of Def. 11.21, cf., e.g., Madarász et al., 2006b, Theorem 3.1.

By the above definition of timelike geodesics, we can express what worldlines of inertial observers are in GR space-times $\langle M, L\rangle$. The definition of photonlike geodesics is analogous, and goes as follows.

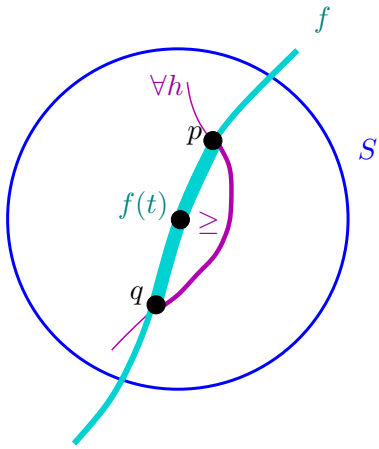


Figure 11.27. f is a timelike geodesic curve.

The curve f is called *photonlike* at p iff the speed of f as seen by the local frame at p equals 1. In more detail, this means that the tangent of the curve $L_p^{-1} \circ f$ at the origin has slope 1. The curve f is called photonlike iff the curve f is photonlike at each of its points p.

We imagine that a photonlike curve can be the worldline of a photon perhaps directed (diverted) by suitably many mirrors.

A *photonlike geodesic* is a photonlike curve f with the property that each point in the curve has a neighborhood in which f is the unique photonlike curve through any two points of f. (In more detail, let F denote the range of f. Then any point in F has a neighborhood S such that whenever f' is a photonlike curve connecting two points of $F \cap S$ and such that F'=*the range of* f' is inside S, we have that $F' \subseteq F$, cf. Hawking and Ellis, 1973, Proposition 4.5.3.)

We imagine that photonlike geodesics are worldlines of photons. Let us notice at this point that a GR space-time $\langle M, L \rangle$ determines "inertial motion" and also determines how photons move.

3.4 The global grid seen with the eyes of the local grids: general relativistic space-time in metric-tensor field form

In the previous subsection, we imported special relativistic notions by using the local frames, the L_p's. We used the inverse L_p^{-1} of the affine transformation L_p to translate our general relativistic "problems", or "questions", to special relativity, and then we used L_p to bring back the answers special relativity gave us. Since we will use the notions of special relativity this way all the time, it is useful to "transport", via L_p, the most useful notions of special relativity

themselves to our general relativistic frame GFR to be ready for use when we need them. Thus, in each point p of GFR, we can "store", e.g., relativistic squared distance $\boldsymbol{\mu}$ of special relativity transported via L_p; we will denote this by $\boldsymbol{\mu}_p$ and we will call it the"local special relativistic squared distance $\boldsymbol{\mu}$ at p". This way we will get "fields" of notions, where we use the word "field" as in analysis and not as in algebra (as mentioned on p. 667). Special relativistic squared distance (i.e. Minkowski distance) $\boldsymbol{\mu}$ and scalar product g_4 were defined on p. 638, p. 656. Here we use the simplified version $\mathsf{g}(p,q) = \mathsf{g}_4(p,\overline{0},q,\overline{0})$ of g as introduced on p. 657.

DEFINITION 11.22 (FIELDS OF "TRANSPORTED" SPECIAL RELATIVITY NOTIONS) *Let $\langle M, L\rangle$ be a general relativistic space-time. For any $p \in M$ we define $\boldsymbol{\mu}_p : \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}$ as follows: for any $q, r \in \mathbb{R}^4$*

$\boldsymbol{\mu}_p(q,r) := \boldsymbol{\mu}(L_p^{-1}(q), L_p^{-1}(r))$; *and similarly we define*

$\mathsf{g}_p(q,r) := \mathsf{g}(L_p^{-1}(q), L_p^{-1}(r))$. *See Fig. 11.28.*

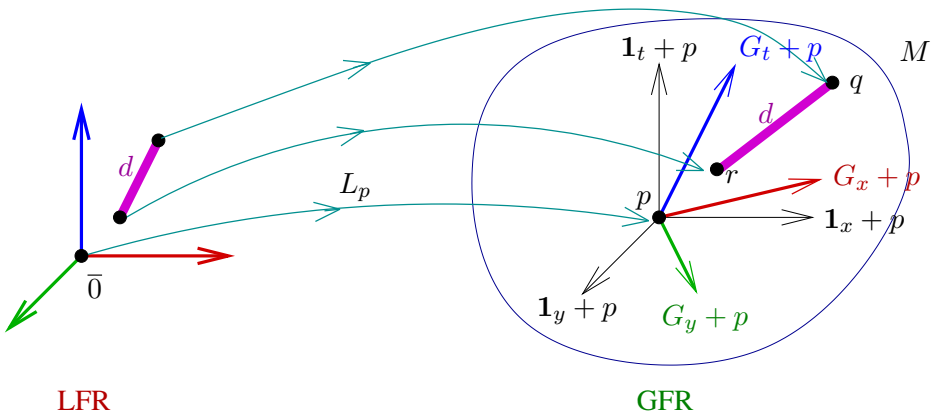


Figure 11.28. The metric $\boldsymbol{\mu}_p$ of local special relativity at p. $d = \boldsymbol{\mu}_p(q,r)$.

In the literature, the most often used form for specifying a GR space-time is by transporting the Minkowski scalar product g to each point p of GFR. The reason for this is that it is easy to make calculations with these data. From $\langle M, L\rangle$ then we get $\langle M, \mathsf{g}_p\rangle_{p\in M}$, usually just written as $\langle M, \overline{\mathsf{g}}\rangle$. We call $\overline{\mathsf{g}} := \langle \mathsf{g}_p : p \in M\rangle$ the *metric-tensor field* of $\langle M, L\rangle$, and we call $\langle M, \overline{\mathsf{g}}\rangle$ the *metric-tensor field form* of $\langle M, L\rangle$.

Since g is linear in its two arguments, the most convenient way of specifying $\mathsf{g}_p : \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}$ is to specify $\mathsf{g}_p(\mathbf{1}_i + p, \mathbf{1}_j + p)$ for all $1 \le i, j \le 4$:

$$\mathsf{g}_{ij}(p) := \mathsf{g}_p(\mathbf{1}_i + p, \mathbf{1}_j + p) := \mathsf{g}(L_p^{-1}(\mathbf{1}_i + p), L_p^{-1}(\mathbf{1}_j + p)), \quad \text{for all } 1 \le i, j \le 4.$$

Then we can specify $\langle M, \overline{\mathsf{g}}\rangle$ by associating the 4 by 4 matrix $(\mathsf{g}_{ij}(p) : 1 \leq i, j \leq 4)$ to each point $p \in M$. What is the meaning of the $\mathsf{g}_{ij}(p)$'s? Well, $\sqrt{|\mathsf{g}_{ii}(p)|}$ tells us how long the i-th unit-vector of the big grid GFR is in the eye of the local special relativity frame at p. On the other hand, $\mathsf{g}_{ij}(p)$ for $i \neq j$ tells us what "angle" between the unit-vectors $\mathbf{1}_i$ and $\mathbf{1}_j$ of the big GFR is as seen by the local special relativity at p. If $\mathsf{g}_{ij}(p) = 0$, then the local special relativity at p (also) thinks that $\mathbf{1}_i$ and $\mathbf{1}_j$ are orthogonal to each other. If $\mathsf{g}_{tt}(p) = 1$, then time at the local special relativity at p flows just as in the GFR grid. If, say, $\mathsf{g}_{tt}(p) = 4$, then two hours pass in the local special relativity at p while in the big grid only one hour passes, hence local special relativity LFR-time at p is twice as fast as coordinate GFR-time. In general, $\mathsf{g}_{tt}(p) > 0$ tells us how the local special relativity sitting at p sees "time of the coordinate grid GFR" to flow (how much slower or faster). If $\mathsf{g}_{tt}(p)$ is negative, then the local special relativity at p "sees" the time-axis of the GFR as a spatial direction, and not as a "temporal direction". This means that in the local special relativity at p, no observer can "move/live" in the space-time direction $\mathbf{1}_t$ of the GFR. If $\mathsf{g}_{xx}(p) = -1$, then the local special relativity at p sees that spatial distance along the x-axis behaves like the one in the big GFR-grid. Also, $\mathsf{g}_{ii}(p) > 0$ iff $\mathbf{1}_i$ of GFR is timelike as seen by the LFR. Etc.

By definition, $\boldsymbol{\mu}_p(q, r)$ is the relativistic squared distance between q and r as seen by the LFR at p, and $\boldsymbol{\mu}_p$ can be expressed using the g_{ij}'s as follows:

$$\boldsymbol{\mu}_p(q, r) = \sum\{\mathsf{g}_{ij}(p) * (q_i - p_i) * (r_j - p_j) : 1 \leq i, j \leq 4\}.$$

The "infinitesimal version" of the above formula is called the *line-element*, $(\star)$ below. In this chapter, we will use the line-element only as an economic linguistic device for specifying the matrix $(\mathsf{g}_{ij}(p) : 1 \leq i, j \leq 4)$. Namely, for $p \in M$, the line-element at p is

$$(\star)\ \ \mathsf{d}s^2(p) = \sum\{a_{ij}\mathsf{d}i\mathsf{d}j : 1 \leq i \leq j \leq 4\}.$$

In the above, we consider $\mathsf{d}s^2$, $\mathsf{d}1, \ldots, \mathsf{d}4$ as "specific linguistic markers", the information content of the line-element $(\star)$ above is

$\mathsf{g}_{ii}(p) = a_{ii}$ for $1 \leq i \leq 4$, and

$\mathsf{g}_{ij}(p) = \mathsf{g}_{ji}(p) = \frac{1}{2}a_{ij}$ for $1 \leq i < j \leq 4$.

E.g. if at $p \in M$ the line-element is $\mathsf{d}s^2(p) = \mathsf{d}t^2 - \mathsf{d}x^2 - \mathsf{d}y^2 - \mathsf{d}z^2$, then $\mathsf{g}_{tt}(p) = 1$, $\mathsf{g}_{xx}(p) = \mathsf{g}_{yy}(p) = \mathsf{g}_{zz}(p) = -1$, and $\mathsf{g}_{ij}(p) = 0$ for $i \neq j$. For more examples see Sec. 4. In Sec. 4 we will use the line-element in specifying a given space-time $\langle M, L\rangle$, not only because of its economy, but also in order to keep comparability with the literature. We want to emphasize that, in this chapter, the line-element is just a convenient linguistic way of specifying g_p, we will not attach independent meanings to $\mathsf{d}s^2$, $\mathsf{d}1, \ldots, \mathsf{d}4$.

Let $\langle M, L\rangle$ be a GR space-time and let $G = \langle G_t, \ldots, G_z\rangle$ and $\langle M, \overline{\mathsf{g}}\rangle$ be its vector-fields and metric-tensor field forms, respectively. Now, G contains the same information as $\langle M, L\rangle$, but $\langle M, \overline{\mathsf{g}}\rangle$ contains slightly less information. However, as we will see in the next section, the really relevant "information" of a space-time $\langle M, L\rangle$ is what is contained in $\langle M, \overline{\mathsf{g}}\rangle$. We use GR space-times in the form $\langle M, L\rangle$ or G because these are easy to draw (visualize), see Figs. 11.30, 11.31, 11.33, 11.40, 11.34; while a GR space-time in the form of $\langle M, \overline{\mathsf{g}}\rangle$ is not so convenient to draw.

From the perspective of the present subsection, in a GR space-time $\langle G_t, \ldots, G_z\rangle$, the tetrad $\langle G_t(p), \ldots, G_z(p)\rangle$ tells us how the big GFR sees the local special relativity unit-vectors of an arbitrarily chosen observer in the local special relativity space-time that "sits" at p. On the other hand, the matrix $(\mathsf{g}_{ij}(p) : 1 \leq i, j \leq 4)$ tells us how the local special relativity space-time sitting at p "sees" the unit-vectors of the big global frame GFR!

3.5 Isomorphisms between general relativistic space-times

We said that the big global frame carries no physical meaning, and only timelike and photonlike geodesics carry physical meanings, everything else (e.g. gravity) can be defined from these geodesics. We give "meaning" to this statement (or claim) in the form of defining what isomorphisms of GR space-times are.

DEFINITION 11.23 *Let $G = \langle M, L\rangle$ and $G' = \langle M', L'\rangle$ be two general relativistic space-times. An* isomorphism *between these two GR space-times is a bijection $Iso : M \to M'$ such that (i)-(iii) below hold.*

(i) *Both Iso and the inverse of Iso are smooth.*

(ii) *Iso preserves timelike geodesics. In more detail, for any curve $f : I \to M$, f is a timelike geodesic in G iff $f \circ Iso$ is a timelike geodesic in G'.*

(iii) *Iso preserves photonlike geodesics (in the above sense).*

We note that we could omit (iii), because one can prove that it follows from (i)–(ii) above.

By using Def. 11.23, it is difficult to check whether a bijection $Iso : M \to M'$ is an isomorphism if we know only L and L' and we did not compute what the geodesics are in $\langle M, L\rangle$ and in $\langle M', L'\rangle$. Below we give an equivalent definition that uses only the "building blocks" L and L' of the general relativistic space-times.

We will use the notion of the differential of a differentiable function.: When $f : \mathbb{R}^4 \to \mathbb{R}^4$ is differentiable, the *differential* of f at $p \in \mathbb{R}^4$ is the affine transformation $\mathsf{D}(f)_p$ which is closest to f at p. Of the properties of a differential we will mostly use that f and $\mathsf{D}(f)_p$ take p to the same point, and they take a

curve g passing through p to tangent curves, i.e. $f(g)$ and $\mathsf{D}(f)_p(g)$ are tangent at $f(p)$.

THEOREM 11.24 (EQUIVALENT FORM OF DEFINITION OF ISOMORPHISMS) *Let $G = \langle M, L\rangle$ and $G' = \langle M', L'\rangle$ be general relativistic space-times and let $\langle M, \overline{\mathsf{g}}\rangle$ and $\langle M', \overline{\mathsf{g}}'\rangle$ be their metric-tensor field forms, respectively. Let $Iso : M \to M'$ be a smooth bijection such that its inverse is also smooth. Then (i)–(iii) below are equivalent.*

(i) *Iso is an isomorphism between G and G' in the sense of Def. 11.23.*

(ii) *For any $p \in M$, the differential of $L'(Iso(p))^{-1} \circ Iso \circ L(p)$ at the origin is a Lorentz transformation (on LFR) perhaps composed with a space-isometry. See Fig. 11.29.*

(iii) *For any $p, q, r \in M$ we have that*

$$\mathsf{g}_p(q, r) = \mathsf{g}'_{Iso(p)}(\mathsf{D}(Iso)_p(q), \mathsf{D}(Iso)_p(r)).$$

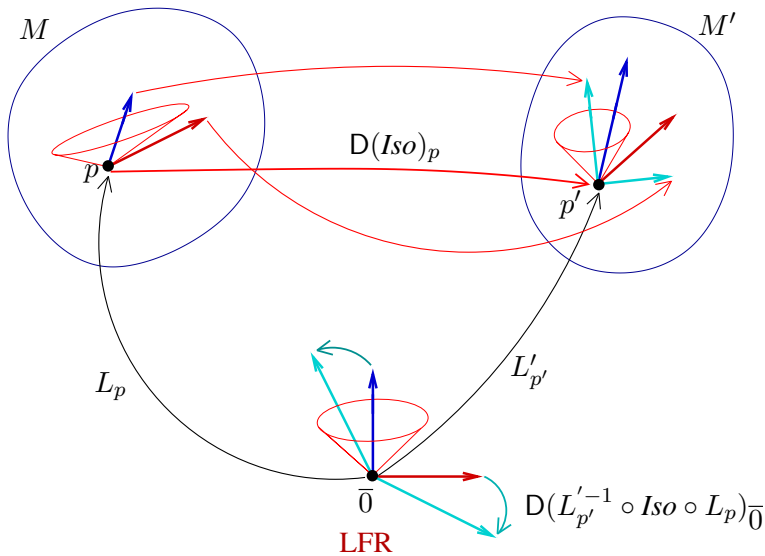


Figure 11.29. Isomorphism between general relativistic space-times.

In connection with Theorem 11.24(ii) above we note the following. Here, the general pattern is $L'(Iso(p))^{-1} \circ Iso \circ L(p) : LFR \to LFR$ because $L(p) : LFR \to M$, $Iso : M \to M'$, and $L'(Iso(p))^{-1} : M' \to LFR$. (More precisely, $L(p) : LFR \to \mathbb{R}^4 \supseteq M$ etc., but that does not matter here.) Recall

that $LFR = \mathbb{R}^4$ refers to the special relativistic frame we are using (in the present section).

By the above, we have the basic building blocks of General Relativity at place and we can start working, we can start discussing black holes, wormholes, the Expanding Universe, etc. So, from here the reader can jump right to Sec. 4 which discusses examples of GR space-times, e.g. black holes. In the next subsection we present a FOL axiomatization of GR space-times which is analogous to **Specrel** and which can serve as a starting point for logic-oriented investigations of GR similar to the ones in Sec. 2.

In Sec. 3.6 we will use the following.

DEFINITION 11.25 *A* Lorentz manifold *is a system of GR space-times connected by commuting partial isomorphisms. I.e. a Lorentz manifold is a system* $\langle\langle G_m, \psi_{mk}\rangle : m, k \in J\rangle$ *such that for all* $m, k \in J$ *the following hold.*

(i) $G_m = \langle M_m, \overline{\mathsf{g}}_m\rangle$ *is the metric-tensor field form of a GR space-time* $\langle M_m, L_m\rangle$ *in the sense of Def. 11.22, except that we do not assume smoothness of* L*, we only assume that* $\overline{\mathsf{g}}_m$ *is smooth.*

(ii) $\langle\psi_{mk} : m, k \in J\rangle$ *is a commuting system of smooth functions with open domains. Commuting means that* $\psi_{kh} \circ \psi_{mk} \subseteq \psi_{mh}$ *for all* $m, k, h \in J$.

(iii) ψ_{mk} *is a partial isomorphism between* G_m *and* G_k*. This means that* ψ_{mk} *is an isomorphism between* G_m *and* G_k *restricted to the domain and the range of* ψ_{mk}*, respectively, in the sense of Theorem 11.24(iii).*

Thus, a Lorentz manifold could be called an "organized system of GR space-times" or a "patched space-time". It is not difficult to show that the above definition of Lorentz manifolds is equivalent with the usual definition in the literature, e.g. with the one in Wald, 1984, pp. 12, 23.

3.6 Axiomatization Genrel of general relativistic space-time in first-order logic

In this subsection we give an axiom system called **Genrel** to general relativistic space-times. This axiom system is formulated in first-order logic and is analogous to **Specrel**, even at the level of the individual axioms.

The vocabulary (or language) of GR space-time models is the same as that in the case of special relativity, with the same intuition as in Sec. 2.1; the motivation for the axioms in **Genrel** is in Secs. 3.2,3.3. In the present subsection we will use arbitrary ordered fields and arbitrary dimensions $n \geq 2$ for the space-time models as in Sec. 2 (and not only $\mathbb{R}$ and $n = 4$). We will use the same notation as in Sec. 2, e.g. we will use ev_m, PLines, etc.

Convention: The elements of Ob are called *inertial observers*. The elements in the domain $\mathsf{Dom}(\mathsf{W}) = \{b : (\exists k, p)\mathsf{W}(b, k, p)\}$ of W are called *(ordinary) observers*.

We will no longer require that an observer m uses the whole coordinate grid Q^n for coordinatizing the events; the part he uses is denoted by $\mathsf{Cd}(m)$, and is defined as $\mathsf{Cd}(m) := \{p \in \mathsf{Q}^n : \mathsf{ev}_m(p) \neq \emptyset\}$. If $p, q \in \bar{t}$, then $[p, q] := \{r \in \bar{t} : p_t \leq r_t \leq q_t\}$.

AxSelf^-: An observer m in his own coordinate system is motionless in the origin, and his worldline is connected, i.e.

$$(\forall m \in \mathsf{B})[\,\mathsf{wline}_m(m) = \bar{t} \cap \mathsf{Cd}(m) \wedge \\ (\forall p, q \in \mathsf{wline}_m(m))[p, q] \subseteq \mathsf{wline}_m(m)\,].$$

We formalize when two subsets h and g of Q^n are tangent at $p \in \mathsf{Q}^n$:

$\mathsf{tangent}(h, g, p)$ means that

$$p \in h \cap g \text{ and } (\forall \varepsilon > 0)(\exists \delta > 0)(\forall s \in [p_t - \delta, p_t + \delta])(\forall q \in h)(\forall r \in g) \\ [\,q_t = r_t = s \;\Rightarrow\; |q - r| \leq \varepsilon * |s - p_t|\,].$$

AxPh^-: An inertial observer m at the origin, where he stands, sees photons move in each direction with speed 1, and each photon meeting m moves with speed 1, i.e.

$$(\forall m \in \mathsf{Ob})[\,(\forall \ell \in \mathsf{PLines})(\forall p \in \ell \cap \mathsf{wline}_m(m))(\exists \mathsf{ph} \in \mathsf{Ph}) \\ \mathsf{tangent}(\ell, \mathsf{wline}_m(\mathsf{ph}), p) \wedge (\forall \mathsf{ph} \in \mathsf{Ph})(\forall p \in \mathsf{wline}_m(\mathsf{ph}) \cap \\ \mathsf{wline}_m(m))(\exists \ell \in \mathsf{PLines})\mathsf{tangent}(\ell, \mathsf{wline}_m(\mathsf{ph}), p)\,].$$

AxThEx^-: An inertial observer m at the origin, where he stands, sees inertial observers move in each direction with speeds < 1, and sees at least one inertial observer in each event, i.e.

$$(\forall m \in \mathsf{Ob})(\forall \ell \in \mathsf{TLines})(\forall p \in \ell \cap \mathsf{wline}_m(m))(\exists k \in \mathsf{Ob}) \\ \mathsf{tangent}(\ell, \mathsf{wline}_m(k), p) \quad \text{and} \quad (\forall p \in \mathsf{Cd}(m))(\exists k \in \mathsf{Ob}) k \in \mathsf{ev}_m(p).$$

$\mathsf{AxSelf}^-, \mathsf{AxPh}^-, \mathsf{AxThEx}^-$ express that an inertial observer experiences special relativity in the space-time location where he is. Next we formalize a generalization of AxLine to general relativity. It will say that in each inertial observer's worldview, the worldlines of inertial observers and photons are timelike and photonlike geodesics, respectively. In formulating this axiom, we will follow the definitions given in Sec. 3.3. We quantified over curves in the definition of geodesics. Since we want to use the language of first-order logic, instead of arbitrary (smooth) curves, we will quantify over bodies representing special curves; namely 3 times continuously differentiable curves which can be defined by first-order logic formulas. This will be the only difference in the definition.

Let us call a curve r-*smooth* if it is r-times continuously differentiable. In general relativity, it is enough to use 3-smooth curves in place of arbitrarily

smooth curves. E.g. for defining curvature, the Riemann-tensor etc., one needs only 3-smooth ingredients in place of smooth ingredients.

Let ψ be a first-order logic formula in our present vocabulary, and assume that the free variables of ψ are among $t, x_1, \ldots, x_n, y_1, \ldots, y_r$ where $t, x_1, \ldots, x_n$ are variables of sort Q. We will denote this assumption as $\psi = \psi(t, \overline{x}, \vec{y}\,)$. We can easily express in first-order logic that, at parameter $\vec{y}$, the formula ψ defines a 3-smooth curve, we will denote this formula by $\mathsf{curve}(\psi)$. To give a flavor for this definition, we start formulating $\mathsf{curve}(\psi)$.

$\mathsf{fn}(\psi)$ denotes the formula $\forall t(\exists \overline{x}\psi(t, \overline{x}, \vec{y}\,) \to \exists! \overline{x}\psi(t, \overline{x}, \vec{y}\,))$. This expresses that ψ defines a (partial) function at parameter $\vec{y}$. When this is the case, we will denote by $\psi(t)$ the value of this function at t. In a similar way, we can express that the domain of the function defined by ψ is an open interval.

$\mathsf{vel}(\psi, t) = \overline{v}$ denotes the formula $\mathsf{fn}(\psi) \wedge (\forall \varepsilon > 0)(\exists \delta > 0)(\forall s \in [t - \delta, t + \delta])(s \neq t \to |[(\psi(s) - \psi(t))/(s - t)] - \overline{v}| < \varepsilon)$. This formula expresses that the velocity vector of the function defined by ψ at t is $\overline{v} \in \mathsf{Q}^n$. Then we can express that the velocity vector changes with t continuously, i.e. the function defined by ψ is 1-smooth. Similarly, we can express that it is 3-smooth. Let $\mathsf{curve}(\psi)$ denote that ψ defines a 3-smooth function, and the domain of this function is an open interval. We note that $\mathsf{vel}(\psi, t)$ is the tangent-vector of the curve ψ at t. The length of this vector depends on the parametrization of the curve.

By the *worldcurve* of observer k in m's worldview we understand the worldline $\mathsf{wline}_m(k)$ parameterized with the wristwatch time of k. We can define this worldcurve by the formula $\gamma_{mk} = \gamma_{mk}(t, \overline{x}) = \gamma(t, \overline{x}, m, k)$ as follows:

$\gamma_{mk} := \gamma(t, \overline{x}, m, k)$ denotes the formula "$\mathsf{w}_{km}(t, \overline{0}) = \overline{x} \wedge \mathsf{W}(k, k, t, \overline{0})$".

Intuitively, γ_{mk} holds for $t, \overline{x}$ iff m sees k present at coordinates $\overline{x}$ such that k's wristwatch shows t when k is present at $\overline{x}$. Below we use $\gamma_{mk}(t)$ as a function of t.

Finally, we express that $\psi = \psi(t, \overline{x}, \vec{y}\,)$ defines a time-faithful curve, or that ψ defines a photonlike curve, respectively as

$\mathsf{timef}(\psi)$ denotes the formula "$\mathsf{curve}(\psi) \wedge \forall t(\exists \overline{x}\psi(t, \overline{x}) \to$
$(\exists k \in \mathsf{Ob})(\exists s \in \mathsf{Q})[\,\psi(t) = \gamma_{mk}(s) \wedge \mathsf{vel}(\psi, t) = \mathsf{vel}(\gamma_{mk}, s)\,])$".

$\mathsf{phot}(\psi)$ denotes the formula "$\mathsf{curve}(\psi) \wedge \forall t(\exists \overline{x}\psi(t, \overline{x}) \to$
$(\exists \mathsf{ph} \in \mathsf{Ph})[\,\psi(t) \in \mathsf{wline}_m(\mathsf{ph}) \wedge \mathsf{tangent}(\exists t\psi, \mathsf{wline}_m(\mathsf{ph}), \psi(t))\,])$"

To be able to quantify conveniently over parametrically defined timelike and photonlike curves, we will use the following axiom schema which is an analogue

of the Comprehension axiom schema in Set Theory. Below, we are thinking of $\psi(t,\overline{x})$ as defining a curve, hence $t \in \mathsf{Dom}\psi$ abbreviates the formula $\exists\overline{x}\psi(t,\overline{x})$, or equivalently, $\exists\overline{x}\psi(t,\overline{x},\vec{y})$. We systematically do not indicate $\vec{y}$ because in $\exists\overline{x}\psi(t,\overline{x},\vec{y})$, the variables in $\vec{y}$ are free variables, they remain free variables while we use ψ in building up new formulas like $\mathsf{Ax}\exists_\psi$ below and eventually in postulating the axioms, all free variables become universally quantified. Hence e.g. $\mathsf{Ax}\exists_\psi$ looks like $\forall\vec{y}\,(\ldots\psi(t,\overline{x},\vec{y})\ldots)$.

In each inertial observer's worldview, the parametrically definable timefaithful curves are worldcurves of (not necessarily inertial) observers; and the photon-like curves are worldlines of bodies. Formally: Let $\psi(t,\overline{x},\vec{y})$ be a formula. Then

$\mathsf{Ax}\exists_\psi$: $(\mathsf{timef}(\psi) \to (\forall m \in \mathsf{Ob})(\exists b \in \mathsf{B})(\forall t \in \mathsf{Dom}\psi)\psi(t) = \gamma_{mb}(t)) \wedge$
$(\mathsf{phot}(\psi) \to (\forall m \in \mathsf{Ob})(\exists b \in \mathsf{B})\{\psi(t) : t \in \mathsf{Dom}\psi\} = \mathsf{wline}_m(b))$.

$\mathsf{COMPR} := \{\mathsf{Ax}\exists_\psi : \psi$ is a formula of our vocabulary$\}$.

For any $p \in \mathsf{Q}^n$ and $\varepsilon > 0$ let $S(p,\varepsilon) := \{q \in \mathsf{Q}^n : |q-p| < \varepsilon\}$, the open ball (or sphere) of radius ε with center p. We now can formulate

"γ_{mk} *is a timelike geodesic*" iff

$\mathsf{timef}(\gamma_{mk}) \wedge (\forall p \in \mathsf{wline}_m(k))(\exists\varepsilon > 0)(\forall q,r \in \mathsf{wline}_m(k) \cap S(p,\varepsilon))$
$(\forall b \in \mathsf{B})[\,\mathsf{timef}(\gamma_{mb}) \wedge \mathsf{wline}_m(b) \subseteq S(p,\varepsilon) \wedge q = \gamma_{mb}(t') = \gamma_{mk}(t) \wedge$
$r = \gamma_{mb}(s') = \gamma_{mk}(s) \quad \Rightarrow \quad |t-s| \geq |t'-s'|\,]$.

"$\mathsf{wline}_m(\mathsf{ph})$ *is a photonlike geodesic*" can be expressed analogously (see p. 671 where photonlike geodesics were defined for $\langle M, L\rangle$).

AxLine^-: In each inertial observer's worldview, the worldlines of inertial observers and photons are geodesics. Formally:

$(\forall m,k \in \mathsf{Ob})$"$\gamma_{mk}$ is a timelike geodesic" and

$(\forall m \in \mathsf{Ob})(\forall \mathsf{ph} \in \mathsf{Ph})$"$\mathsf{wline}_m(\mathsf{ph})$ is a photonlike geodesic".

Now we turn to formulating a generalization of $\mathsf{AxEvent}$. It expresses that if m observes k participate in an event, then k himself "sees" this event. Further, if k sees an event that m sees, then k sees all events which occur "near" this event in m's worldview.

$\mathsf{AxEvent}^-$:

$(\forall m,k \in \mathsf{Ob})(\forall p \in \mathsf{Q}^n)(k \in \mathsf{ev}_m(p) \;\Rightarrow\; (\exists q \in \mathsf{Q}^n)\mathsf{ev}_k(q) = \mathsf{ev}_m(p)) \wedge$
$(\mathsf{Dom}(\mathsf{w}_{mk})$ is open and w_{mk} is a 3-smooth function$))$.

To formulate a generalization of AxSim, we will use a variant of AxSim that works for $n = 2$, too. For more on this variant we refer to Andréka et al., 2002, Secs. 2.8, 3.9, Andréka et al., 2006b, p.162, Andréka et al., 1999.

AxSim^-: Any two inertial observers see each other's wristwatches run slow with the same ratio when they meet:

$$(\forall m, k \in \mathsf{Ob})(\forall t, s \in \mathsf{Q})[\,\mathsf{ev}_m(\gamma_{mk}(t)) = \mathsf{ev}_k(\gamma_{km}(s)) \;\Rightarrow \; |\mathsf{vel}(\gamma_{mk}, t)| = |\mathsf{vel}(\gamma_{km}, s)|\,].$$

To be able to use the notions of continuity and differentiability etc. in arbitrary fields in place of $\mathbb{R}$ properly, we need the axiom schema of Continuity. Reasons and details for this can be found e.g. in Madarász et al., 2006b, Goldblatt, 1987, van Benthem, 1983, p. 29.

AxSup_ψ**:** is a formula expressing that every subset of Q defined by $\psi(t, \vec{y})$ with parameter $\vec{y}$ has a supremum if it is non-empty and bounded. Formally AxSup_ψ is: $\exists t' \forall t[\psi(t, \overline{y}) \to t < t'] \to \exists t'(\forall t[\psi(t, \overline{y}) \to t < t'] \wedge \forall t''[\forall t(\psi(t, \overline{y}) \to t < t'') \to t'' \geq t'])$.

CONT**:** $:= \{\mathsf{AxSup}_\psi : \psi$ is a formula in our vocabulary$\}$.

The formula schema CONT above is a variant of Tarski's first-order logic version of Hilbert's axiom of continuity in his axiomatization of Euclidean geometry. It is also strongly related to the induction axiom schema in the dynamic logic of actions in the sense of Sain, 1986, Andréka et al., 1982.

$$\textbf{Genrel} := \{\mathsf{AxSelf}^-, \mathsf{AxLine}^-, \mathsf{AxThEx}^-, \mathsf{AxPh}^-, \mathsf{AxEvent}^-, \mathsf{AxSim}^-\} \cup \{\mathsf{AxField}\} \cup \mathsf{CONT} \cup \mathsf{COMPR}.$$

As a first theorem we state that special relativity is the special case of general relativity where the worldlines of all observers and photons are straight lines; and where the Light Axiom holds. This is stated in the next theorem. We conjecture that a much weaker axiom suffices in place of the Light Axiom.

THEOREM 11.26 **Genrel**$\cup\{\mathsf{AxLine}, \mathsf{AxPh}\}$ *is equivalent to* **Specrel**$\cup\mathsf{CONT}$ $\cup\mathsf{COMPR}$ *in the sense that they have the same models, if* $n \geq 3$.

Proof outline Assume **Genrel** + AxLine + AxPh. First one proves that $(\forall m \in \mathsf{Ob})\mathsf{Cd}(m) = \mathsf{Q}^n$, by AxLine. Then $\mathsf{AxEvent}$ can be proved from $\mathsf{AxLine}, \mathsf{AxPh}$ and $\mathsf{AxEvent}^-$ along the lines of the proof in Andréka et al., 2002, pp. 98-100. Now the axiom AxEOb used in Madarász et al., 2004 holds, so one gets that the worldview transformations are affine mappings, by Madarász et al., 2004, Theorem 1. It is proved in Andréka et al., 2002, Theorem 3.9.11 that AxSim^- implies AxSim when the worldview transformations are affine. This proves one direction of Theorem 11.26. In the other direction we have to prove in an axiomatic setting that timelike and photonlike straight lines are timelike and photonlike geodesics, respectively. This is done, basically, in Madarász et al., 2006b, Theorem 3.1. QED

Next we show that the metric-geometric forms of the models of **Genrel** are *Lorentz manifolds* and each Lorentz manifold is the metric-geometric form of a model of **Genrel**. This will be the analogue of Theorem 11.16 in Sec. 2.6.

DEFINITION 11.27 *Let $\mathfrak{Q} = \langle \mathsf{Q}, +, *, \leq \rangle$ be an ordered quadratic field. By a* 3-smooth n-dimensional *Lorentz manifold* over $\mathfrak{Q}$ *we understand* $\langle \langle M_m, \overline{\mathsf{g}}_m, \psi_{mk}, \mathfrak{Q} \rangle : m, k \in J \rangle$ *where* $\langle \langle \langle M_m, \overline{\mathsf{g}}_m \rangle, \psi_{mk} \rangle : m, k \in J \rangle$ *is a Lorentz manifold in the sense of Def. 11.25 with the following changes:*

(a) *in place of* $\mathbb{R}^4$ *we use* $\mathfrak{Q}^n$*, and*

(b) *in place of requiring the metric-tensor fields* g_m *and the partial isomorphisms* ψ_{mk} *to be smooth, we only require that they be 3-smooth.*

The following theorem says that the models of **Genrel** are exactly the 3-smooth Lorentz manifolds over ordered real-closed fields, up to ignoring some "decorations" on the **Genrel** side. Theorem 11.28 can be considered as a completeness theorem for **Genrel**.

THEOREM 11.28 (COMPLETENESS THEOREM FOR **Genrel**) *Assume* $n \geq 3$.

(i) *There is a theory* $\mathbf{Comp}^-$ *analogous to* **Comp** *such that* $\mathbf{Genrel} \cup \mathbf{Comp}^-$ *is definitionally equivalent with the class* LM *of all 3-smooth Lorentz manifolds over ordered real-closed fields.*

(ii) *There is a definable function* Lm *that maps the class of all models of* **Genrel** *onto the class* LM*. Moreover, if* $\mathfrak{M} \models \mathbf{Catrel}$ *then* $\mathsf{Lm}(\mathfrak{M})$ *is definitionally equivalent with the Minkowski geometry* $\mathsf{Mg}(\mathfrak{M})$ *of* $\mathfrak{M}$.

The proof of Theorem 11.28 is based on the following proposition, which states that the models of **Genrel** are "locally special relativistic".

PROPOSITION 11.29 *Assume that* $\mathfrak{M} \models \mathbf{Genrel}$ *and* $m, k \in \mathsf{Ob}$*. If* $p \in \mathsf{wline}_m(k) \cap \bar{t}$ *then* $\mathsf{D}(\mathsf{w}_{mk})_p$ *exists and it preserves relativistic (Minkowski) distance* $\boldsymbol{\mu}$.

Proof outline Assume $p \in \mathsf{wline}_m(k) \cap \bar{t}$. Then $k, m \in \mathsf{ev}_m(p)$ by AxSelf^-, and so $p \in \mathsf{Dom}(\mathsf{w}_{mk})$ and $F := \mathsf{D}(\mathsf{w}_{mk})_p$ exists by $\mathsf{AxEvent}^-$. By AxPh^-, and the definition of w_{mk}, F takes PLines and only PLines going through p to PLines going through $p' := \mathsf{w}_{mk}(p)$. Thus F is a bijection and by AxSim^- it preserves $\boldsymbol{\mu}$. QED

To give an idea for the proof of Theorem 11.28, we define $\mathsf{Lm}(\mathfrak{M})$ for $\mathfrak{M} \models \mathbf{Genrel}$. Let us fix a model $\mathfrak{M} = \langle \mathsf{Q}, +, *, \leq; \mathsf{B}, \mathsf{Ob}, \mathsf{Ph}; \mathsf{W} \rangle \models \mathbf{Genrel}$

and let $\mathfrak{Q} = \langle \mathsf{Q}, +, *, \leq \rangle$. Let $m \in \mathsf{Ob}$ and $p \in \mathsf{Cd}(m)$. Let $k \in \mathsf{ev}_m(p) \cap \mathsf{Ob}$ be arbitrary (such a k exists by AxThEx^-) and let $p' := \mathsf{w}_{mk}(p)$. We define

$L_{p,k} := \mathsf{D}(\mathsf{w}_{mk})_{p'} \circ \tau(p')$, where $\tau(p') : \mathsf{Q}^n \to \mathsf{Q}^n$ is "translation with p'", and

$\mathsf{g}_{p,k}$ is the metric-tensor field belonging to this $L_{p,k}$ as defined in Def. 11.22, i.e. $\mathsf{g}_{p,k}(q,r) := \mathsf{g}(L_{p,k}^{-1}(q), L_{p,k}^{-1}(r))$, for all $q, r \in \mathsf{Q}^n$.

It follows from Proposition 11.29 that, though $L_{p,k}$ depends on how we choose $k \in \mathsf{ev}_m(p)$, the metric-tensor $\mathsf{g}_{p,k}$ does not depend on how we choose $k \in \mathsf{ev}_m(p)$. Therefore we will omit the index k from the notation:

$$\mathsf{g}_p := \mathsf{g}_{p,k}, \quad \overline{\mathsf{g}}_m := \langle \mathsf{g}_p : p \in \mathsf{Cd}(m) \rangle, \quad G_m := \langle \mathsf{Cd}(m), \overline{\mathsf{g}}_m \rangle, \quad \text{and}$$

$$\mathsf{Lm}(\mathfrak{M}) := \langle \langle \mathsf{Cd}(m), \overline{\mathsf{g}}_m, \mathsf{w}_{mk}, \mathfrak{Q} \rangle : m, k \in \mathsf{Ob} \rangle.$$

CLAIM 11.30 *Assume* $\mathfrak{M} \models$ **Genrel**. *The following (i),(ii) hold.*

(i) $\mathsf{Lm}(\mathfrak{M})$ *is a 3-smooth Lorentz manifold over* $\mathfrak{Q}$*, and* $\mathfrak{Q}$ *is an ordered real-closed field.*

(ii) *The worldlines* $\mathsf{wline}_m(k), \mathsf{wline}_m(\mathsf{ph})$ *of observers and photons in* m*'s worldview are timelike and photonlike geodesics in* G_m*. Conversely, any timelike geodesic* ℓ *is a worldline of an observer locally, i.e.* $(\forall p \in \ell)(\exists \varepsilon > 0)(\exists k \in \mathsf{Ob})\ell \cap S(p, \varepsilon) \subseteq \mathsf{wline}_m(k)$*. The same converse holds for photonlike geodesics, too.*

We outlined above that the geometry of each model of **Genrel** is a Lorentz 3-smooth manifold. The converse is also true, each Lorentz 3-smooth manifold over an ordered real-closed field is isomorphic (as a manifold) to the geometry $\mathsf{Lm}(\mathfrak{M})$ of a model $\mathfrak{M}$ of **Genrel**.

The extension $\mathbf{Comp}^-$ for **Genrel** is completely analogous with **Comp**. A typical axiom in $\mathbf{Comp}^-$ states that if the worldlines of two photons ph_1 and ph_2 coincide for all observers, then $\mathsf{ph}_1 = \mathsf{ph}_2$. Further, the worldline of a photon is a maximal photonlike geodesic. The point in the axioms in $\mathbf{Comp}^-$ is to eliminate things like the multiplicity of otherwise undistinguishable objects (like ph_1 and ph_2 above) which cannot be defined over LM because they are undistinguishable. In some sense these statements are incarnations of Occam's razor.

We omit the rest of the idea for proof of Theorem 11.26. QED

What we call the worldview of an inertial observer $m \in \mathsf{Ob}$ in **Genrel** corresponds to "spaceships with ship-drive switched off": the worldline of the

center of mass is a geodesic, but we did not care about whether the spaceship rotates or not. One can base an axiomatization of GR on worldviews of the so-called "*local inertial frames*" (LIF's, cf. Rindler, 2001, pp. 177–179) which correspond to nonrotating inertial spaceships. LIF's reflect the local special relativity more closely (e.g. they do not rotate). However, LIF's can be defined in models of **Genrel** and the axiomatization based on LIF's would provide the same geometrical entities $\mathsf{Lm}(\mathfrak{M})$ behind the models as the present **Genrel** does. The role of Einstein's field equations in interpreting **Genrel** is touched upon in Sec. 4.5.

4. Black holes, wormholes, timewarp. Distinguished general relativistic space-times

In Sec. 3.6 we introduced the first-order logic theory **Genrel** of general relativity. In such a situation, the next natural thing to do is to turn to the intended models of the theory in question, and to discuss what these models look like. Indeed, we will do this in the present section, we will study some of the intended models of the theory **Genrel**. A difference between special relativity and GR is that while the special theory had basically one intended model (namely Minkowski space-time), the general theory has many nonisomorphic intended models, as we will see below.

It will be convenient for us to study the models of **Genrel** in their geometric forms. Hence we will speak about GR space-times $\langle M, L\rangle$, but it is important to remember that in Sec. 3.6 we saw that such a space-time $\langle M, L\rangle$ is equivalent with a **Genrel** model $\mathfrak{M} \models$ **Genrel**. So each one of the GR space-times in this section represents a distinguished **Genrel** model, and discussing these will shed some light on the semantic aspects of **Genrel**.

For discussing the models of **Genrel** we leave the realm of first-order logic and then we work in mathematics proper, the reason for which is that by Tarski's theorem one cannot satisfactorily describe the semantics of a language $\mathcal{L}$ inside the framework of $\mathcal{L}$ itself. To study the semantics of $\mathcal{L}$, we have to rise above $\mathcal{L}$ and use the meta-language of $\mathcal{L}$. In our case, this metalanguage is ordinary mathematics (or equivalently Set Theory, say ZF).

4.1 Special relativity as special case of general relativity

Minkowski space-time is $\mathsf{G}_{sr} = \langle M, L\rangle$ where $M = \mathbb{R}^4$ and $L(p)$ is translation with p (i.e. $L_p(q) = p + q$ for any $q \in \mathbb{R}^4$), for all $p \in \mathbb{R}^4$. In vector-fields form this is $\langle G_1, \ldots, G_4\rangle$ where $G_i(p) = \mathbf{1}_i$ for all $p \in \mathbb{R}^4$ and $1 \leq i \leq 4$, see upper part of Fig. 11.30. In metric-tensor field form Minkowski space-time is

$\langle M, \overline{\mathsf{g}}\rangle$ where the 4 by 4 matrix $(\mathsf{g}_{ij}(p) : 1 \leq i, j \leq 4)$ at $p \in \mathbb{R}^4$ is

$$\mathsf{g}(p) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

i.e. the line-element is $\mathsf{d}s^2 = \mathsf{d}t^2 - \mathsf{d}x^2 - \mathsf{d}y^2 - \mathsf{d}z^2$. The timelike and photonlike geodesics in this space-time are the straight lines of slope < 1 and of slope 1, respectively. The automorphisms (i.e. isomorphisms onto itself) of Minkowski space-time G_{sr} are exactly the possible worldview transformations in a model of **Specrel** (where $\langle \mathsf{Q}, \ldots\rangle = \langle \mathbb{R}, \ldots\rangle$), cf. Theorem 11.10.

Worldview of a uniformly accelerated observer in Specrel:

We let $n = 2$, for simplicity. Consider the following space-time $\mathsf{G}_{ua} = \langle M, L\rangle$ where $M = \{\langle t, x\rangle \in \mathbb{R}^2 : x > 0\}$ and $G_t(p) = \langle 1/p_2, 0\rangle$, $G_x(p) = \langle 0, 1\rangle$ for all $p = \langle p_1, p_2\rangle \in M$, cf. the lower part of Fig. 11.30. Thus the line-element is

$\mathsf{d}s^2 = x^2\mathsf{d}t^2 - \mathsf{d}x^2$, and the g_{ij}-matrix is

$$\mathsf{g}(t, x) = \begin{pmatrix} x^2 & 0 \\ 0 & -1 \end{pmatrix}.$$

In this space-time, as we approach the origin, local LFR clocks tick slower and slower beyond any limit compared with coordinate GFR time, and local LFR clocks tick faster and faster beyond any limit compared with coordinate GFR time as we move away from the origin. On the other hand, local meter-rods do not change along the x-axis, local LFR spatial distances agree with coordinate GFR spatial distances. This space-time looks different from Minkowski space-time G_{sr}, but in fact it is isomorphic to a sub-space-time of 2-dimensional G_{sr}. The isomorphism denoted by *Iso* is represented in Fig. 11.30, it maps M of G_{ua} bijectively onto $\{\langle t, x\rangle : |t| < x\}$. The space-time G_{ua} is the worldview (or space-time) of a uniformly accelerated observer k who lives in Minkowski space-time, with uniform (relativistic) acceleration $a = 1$. (The space-time for arbitrary uniform acceleration a is given by $G_t(p) = \langle 1/(ap_2), 0\rangle$, $G_x(p) = \langle 0, 1\rangle$.) One can think of G_{sr} as the worldview of an inertial observer m in special relativity, and then *Iso* is the worldview transformation w_{km} between the worldview G_{ua} of accelerated k and the worldview G_{sr} of m.

G_{ua} is rather similar to the exterior of the (2-dimensional tr-slice of the) simplest black hole G_{sb} below, which, in turn, is no longer partially isomorphic to any open part of G_{sr}. Studying the simple space-time G_{ua} of accelerated observers can lead to a deeper understanding of the space-time G_{sb} of the important Schwarzschild black hole to which we turn now. This connection is elaborated e.g. in the textbook Rindler, 2001, Secs. 3.7, 12.4.

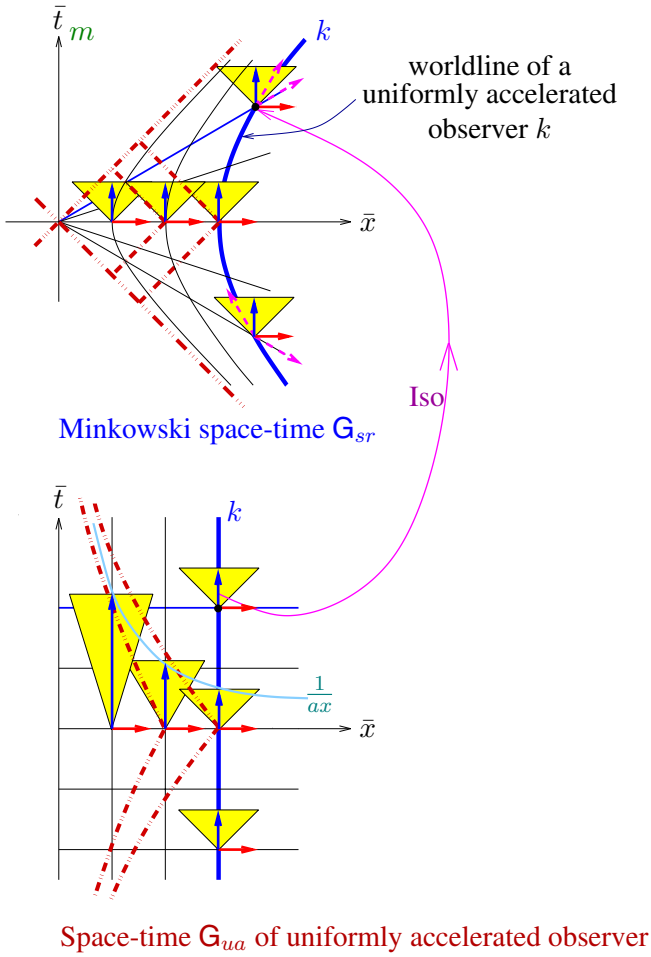


Figure 11.30. Isomorphism from the worldview of a uniformly accelerated observer to Minkowski space-time. (The example in the text is given for uniform acceleration $a = 1$, the figure is for $a = 1/3$.) $Iso(t, x)$, for $t > 0$, is the point p on the Minkowski-circle $M(\bar{0}, x)$ of radius x and with center the origin such that the relativistic arc-length of $M(\bar{0}, x)$ from $\langle 0, x\rangle$ to p is axt. $\frac{1}{ax}$ is the length of G_t at x.

4.2 The Schwarzschild black hole

There is now overwhelming observational evidence for the existence of large black holes in our universe. In the last 15 years astronomers have observed them. At the same time, black holes have really fantastic properties. Black hole physics is at the cutting edge of modern science.

There are many kinds of black holes, the Schwarzschild black hole is the simplest one. We will recall more exotic ones in Sec. 4.3. Why is the Schwarzschild black hole important? Here are four reasons for this: it is the simplest form of relativistic gravity (all the mass is in one point), it is an idealization of the gravitational space-time of our own Sun, it is a typical general relativistic space-time, and many other GR space-times build on it.

Worldview of a suspended observer far away from the black hole:

Schwarzschild (black hole) space-time is $\mathsf{G}_{sb} := \langle M, L\rangle$ where $M = \{p \in \mathbb{R}^4 : |\langle p_2, p_3, p_4\rangle| \neq 0, 1\}$ and L is given as follows. For any $p = \langle p_1, p_2, p_3, p_4\rangle \in \mathbb{R}^4$ let $\mathbf{r} := \mathbf{r}(p) := \langle 0, p_2, p_3, p_4\rangle$, $r := |\mathbf{r}|$, and $\mathbf{1}_r := \mathbf{1}_r(p) := \frac{1}{r} * \mathbf{r}$. Here, "r" stands for "radius". Now L is specified by the following four vector-fields

For $r > 1$:

$G_t(p) = \sqrt{\frac{r}{r-1}} * \mathbf{1}_t$, $G_x(p) = \sqrt{\frac{r-1}{r}} * \mathbf{1}_r$, the lengths of $G_y(p)$ and $G_z(p)$ are 1, and $G_t(p), G_x(p), G_y(p), G_z(p)$ are pairwise orthogonal.

For $r < 1$:

$G_t(p) = \sqrt{\frac{1-r}{r}} * \mathbf{1}_r$, $G_x(p) = \sqrt{\frac{r}{1-r}} * \mathbf{1}_t$, the lengths of $G_y(p)$ and $G_z(p)$ are 1, and $G_t(p), G_x(p), G_y(p), G_z(p)$ are pairwise orthogonal.

See Fig. 11.31.

G_{sb} in metric-tensor form is the following. We use cylindric-polar coordinates because they are more convenient (by spatial spherical symmetry of the space-time). The line-element is

$$\mathsf{d}s^2 = (1 - \tfrac{1}{r})\mathsf{d}t^2 - (1 - \tfrac{1}{r})^{-1}\mathsf{d}r^2 - r^2\mathsf{d}\varphi^2,$$

where $\mathsf{d}\varphi$ represents two coordinates the usual Euclidean way. Namely, φ represents "space-angle", i.e. φ is the usual Euclidean combination of two polar coordinates θ (longitude) and η (latitude). Formally, $\mathsf{d}\varphi^2 = \mathsf{d}\theta^2 + \sin(\theta)^2\mathsf{d}\eta^2$ (metric on Euclidean unit 2-sphere). We note that by defining the line-element we also defined the metric-tensor field $\overline{\mathsf{g}}$ of Schwarzschild space-time G_{sb}.

In the general form of Schwarzschild space-time, there is a parameter that we chose to be 1 in the above. Namely, the general form of the line-element for the Schwarzschild black hole is

$$\mathsf{d}s^2 = (1 - \tfrac{\mathsf{M}}{r})\mathsf{d}t^2 - (1 - \tfrac{\mathsf{M}}{r})^{-1}\mathsf{d}r^2 - r^2\mathsf{d}\varphi^2,$$

the parameter $\mathsf{M} \in \mathbb{R}$ is thought of as the "mass" of the black hole. M is also called the *radius* or *size* of the black hole. (For historical reasons $2m$ is used for M in the literature, but this does not matter when one wants to understand the "logic" of the black holes.) Similarity with the accelerated observer can be discovered by choosing $x = r - 1$. Then the accelerated line-element (i.e. that of G_{ua}) becomes $\mathsf{d}s^2 = (r-1)^2\mathsf{d}t^2 - \mathsf{d}r^2$. For lack of space we do not discuss this more here, but we note that there is more in this direction in Andréka et al., 2006a.

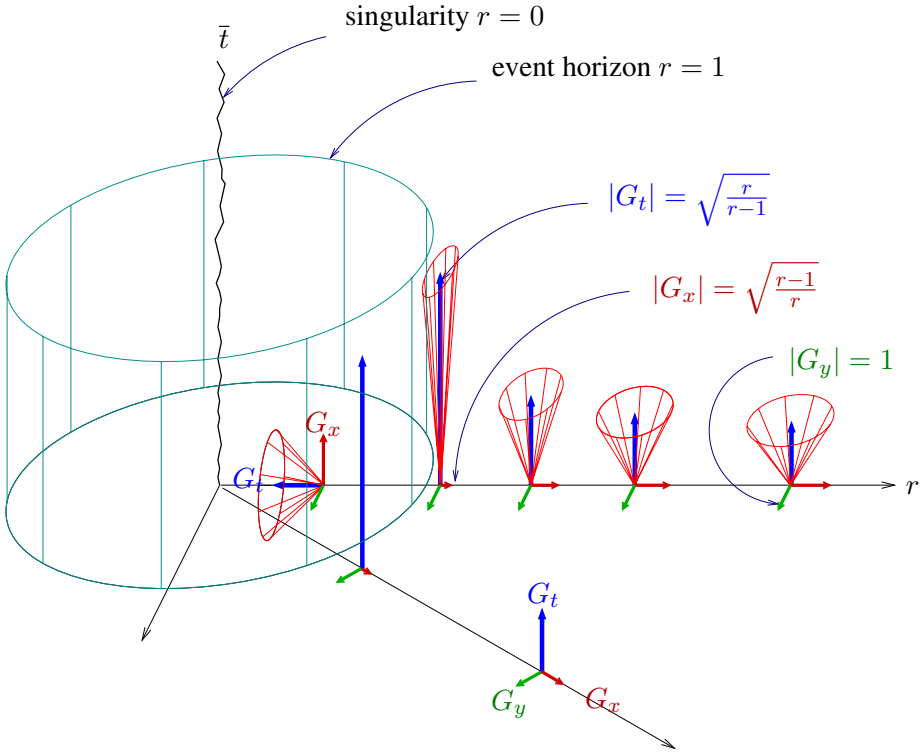


Figure 11.31. Illustration for Schwarzschild space-time. G_t gets longer as we approach $r = 1$ from above, i.e. local time runs slower and slower as we approach $r = 1$. G_x gets shorter as we approach $r = 1$, i.e. there is more and more space (compared to "coordinate-space") as we approach $r = 1$. The length of G_y stays 1, this means that spatial distances orthogonal to the radius agree with coordinate-distances. Time and (radial) space are interchanged in the interior of the black hole, this means that, in the interior, the r-axis is the worldline of an observer, but lines parallel to the time-axis are not possible worldlines. The singularity is in the future of an observer inside the black hole.

The set of coordinate points $p \in \mathbb{R}^4$ with $r = 0$ is called the *singularity* (this coincides with the time-axis), and the set of coordinate points $p \in \mathbb{R}^4$ with $r = 1$ is called the *event horizon* (EH). These are disjoint from the domain of G_{sb}, i.e. we did not associate local clocks and meter-rods to these points. Thinking in terms of the global coordinate frame, the event horizon is a sphere of radius 1 (or M in the more general case), and the singularity is the center of this sphere. Loosely speaking, we will refer to the EH and its interior as the *black hole* (BH). When we want to be more careful, we will refer to the part outside of the EH as the *exterior* of the BH and we will refer to the inside part of the EH as the *interior* of the BH. Later we will see that the interior and the exterior of the BH behave, in some sense, like two different universes connected

by a one-way membrane, namely by the EH. Here one-way membrane means that an observer may fall through the EH into the interior of the BH, but nothing, not even light can come out from the interior. This effect is not yet clear from our present space-time diagram (Fig. 11.32) but it will be clear after we apply the so-called Eddington-Finkelstein re-coordinatization (and extension) to it, cf. Fig. 11.35.

Let us think for a while in terms of the global coordinate system, and let us see what the exterior of the BH looks like. Infinitely far away everything is normal, the farther away we are from the EH, the more "normal life is", e.g. both local time and local meter-rods agree with the global GFR coordinate ones asymptotically. Space-times with this property are called *asymptotically flat*. Asymptotically flat means asymptotically Minkowski (or asymptotically special relativistic), namely as we move away from the BH, space-time becomes more and more like Minkowski space-time is. This can be formalized by saying that as $r(p)$ tends to infinity, so the metric-tensor $\mathsf{g}(p)$ tends to "Minkowski g".

Convention: by coordinate properties (e.g. coordinate time) we always mean the global, GFR-coordinate properties.

Let us now approach the EH from far away. As we get close, local clocks begin to tick radically slower (compared to the global coordinate time), beyond any limit; so metaphorically local clocks "stop" or freeze at the EH. (This is only metaphorical because we did not associate local clocks and meter-rods to the points of the EH. However, this will be helped after the upcoming Eddington-Finkelstein re-coordinatization and then time will really freeze on the EH.) At the same time, local meter-rods in the radial direction get smaller (beyond any limit) as we approach the EH, but local meter-rods orthogonal to the radius of the EH continue to agree with the coordinate meter-rods.

Far away from the BH, GFR-coordinate-speed of light tends to be the same, namely 1, in all directions, but as we approach the EH, the coordinate-speed of light in the radial direction gets radically smaller compared to the coordinate-speed of light in directions orthogonal to the radius (this coordinate "anisotropy" is the reason why the light-cones in Fig. 11.31 close to the EH but in the exterior get "flattened out"); and as we get closer to the EH, the coordinate-speed of light tends to 0 in all directions. We note that the fact that $\mathsf{g}_{ti}(p) = 0$ for $i \neq t$ everywhere means that at each event, the coordinate-speed of light in a spatial direction d and in its opposite direction $-d$ is the same, if measured by the global frame. The above means that the so-called light-cones get infinitely narrow towards the EH but they do not get tilted as seen by the GFR, cf. Fig. 11.31, Fig. 11.32.

Since a timelike curve must stay inside the local light-cones, this means that observers stay all the coordinate time outside the EH, they never "enter" the EH, as seen via the global coordinate grid. We will see that this is only an "artifact" of Schwarzschild's particular choice of global coordinate system (GFR), similar to

the "artifact" mentioned earlier and also in Fig. 11.30. Avoiding of this artifact will be done via the so-called Eddington-Finkelstein re-coordinatization which will be presented soon.

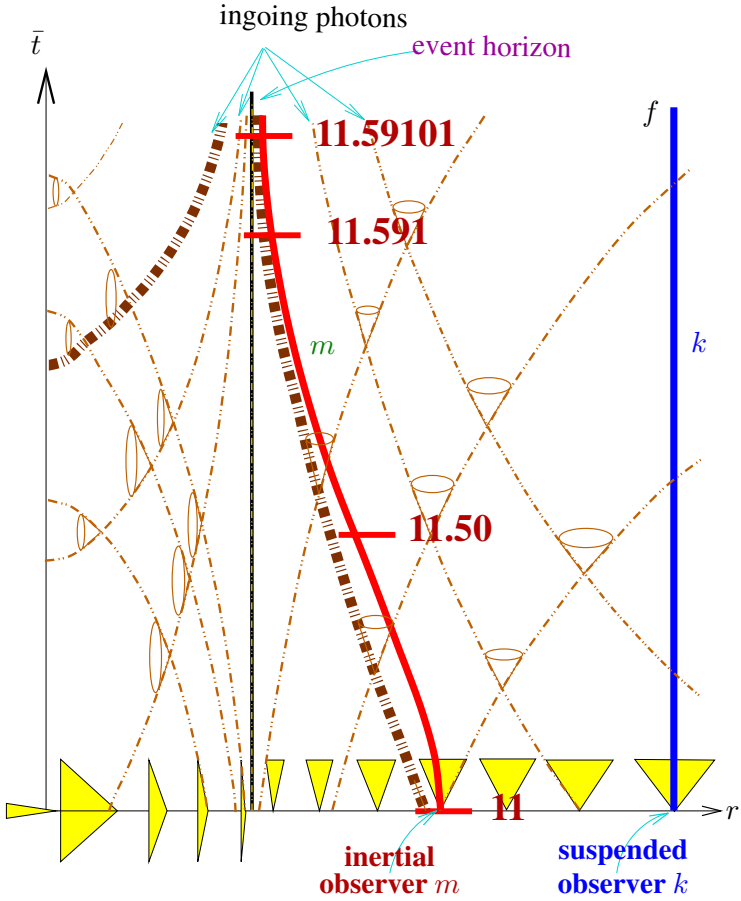


Figure 11.32. The "tr-slice" of a Schwarzschild black hole. This is a geodesically closed 2-dimensional sub-space-time of G_{sb}. The worldline of m is a geodesic, it "bulges" outward because m can maximize his time by bulging outwards. The worldline of k is not a geodesic (because it does not "bulge" outwards). Photons and inertial observers moving in radially "freeze" to the EH. The wristwatch of an in-falling inertial observer slows down "infinitely", and will show times which are bounded.

Let us see what the worldlines of observers and photons in Schwarzschild space-time look like in the exterior of the EH.

Consider, for an example, the curve $f : \mathbb{R} \to \mathbb{R}^4$ where $f(\tau) = \langle \sqrt{1.5} * \tau, 3, 0, 0 \rangle$ for all $\tau \in \mathbb{R}$. (This f is represented in Fig. 11.32 as "suspended

observer".) This is a time-faithful curve, so we can imagine an observer k whose worldcurve this f is; the worldline of k (in G_{sb}) is the vertical line going through $\langle 0, 3, 0, 0\rangle$ and k's wristwatch shows τ at the coordinate point $\langle \sqrt{1.5} * \tau, 3, 0, 0\rangle$. Is this curve f a geodesic? Local clocks at coordinate points farther away from the EH tick faster, so one can "gain time" by moving outwards a little, clocking up a lot of wristwatch time while out there, and then coming back. Special relativistic time-dilation dampens this effect somewhat since the clocks of a fast moving observer slow down. But it is not hard to show that by moving outwards and then coming back with small velocities all the time, one gains time, and there is an optimum "outward-bulging" with maximum gain of time. Hence, vertical lines are not geodesics and the timelike geodesics always "bulge a little outwards", i.e. they accelerate (or "turn"), as seen in GFR, towards the BH. So, f is not a geodesic.

Radial timelike geodesics are similar to the worldlines of pebbles thrown up into the air here on the Earth; with "upward" replaced by "outward"; a geodesic curve which begins to move away from the EH looses (coordinate) speed, eventually it stops and "falls back" towards the EH. According to GR, things thrown up fall back not because gravity of the Earth pulls them with myriad small invisible "hands", but because time ticks slower near the Earth, and faster farther away from the Earth. Newton's apple falls from the tree because of the "gravitational time-dilation" (known also as gravitational red-shift)! This is the first explanation for gravity since its behavior was described by Newton.

There is a similar reason for a photonlike geodesic with a velocity in nonradial direction to bend toward the EH as if the EH "pulled" the geodesic towards itself. So, even photons "fall" (gravitational "light-bending" effect).

Since the BH attracts, if an observer wants to stay at a constant distance from the EH, he has to use fuel for accelerating away from it. We call an observer like k above a "suspended observer". *Suspended* observer means that the worldline of the observer is parallel with the time-axis in the present Schwarzschild GFR coordinate system. However, the notion of "suspended observer" is coordinatization-independent, i.e. observer-independent, because these "vertical" worldlines can be defined by a first-order logic formula in the language of local clocks and meter-rods (i.e. in the language of **Genrel** in Sec. 3.6). In other words, there is an experiment with which an observer can check whether he is suspended or not. In still other words, being suspended is an observational concept.

Let us consider a (timelike or photonlike) geodesic curve that starts out towards the EH in a radial direction. By cylindrical symmetry of the space-time, there is "no reason" for the geodesic to bend right or left, since "right" and "left" are completely alike by symmetry. This implies that a geodesic curve which starts in a radial direction, will always move in this radial direction, i.e.

the (range of the) curve will be a subset of a tr-plane. Thus a tr-plane is a *geodesically closed sub-space-time* of G_{sb}.

Let us review briefly the interior of the EH. In the interior of the EH, time and (radial) space are interchanged, the r-axis is in fact the worldline of a possible inertial observer. Hence the global r-axis is the local time-axis for some LFR "living" in the interior of the BH. Similarly, the $\bar{t}$-axis of the GFR is a spatial direction for this LFR. The singularity is no more a "place" for this observer (or for any observer in the interior), instead, it is like a future time instance like the Big Crunch in usual cosmology, something that will happen in the distant future but not "present" in the "now" of the in-falling observer. Similarly, for this inertial observer inside the BH, the EH is like a time instance or a "time-slice of space-time" which happened sometime in the past like the Big Bang in cosmology. We see the light coming from the Big Bang or from the EH in our past but we cannot influence it causally because it is in our past. Local time at the EH (in the interior) is "infinitely fast" compared to coordinate-time and local meter-rods in the $\bar{t}$-direction are "infinitely long". As we move towards the singularity, local time "slows down" (compared to GFR) beyond any limit, and local meter-rods in the $\bar{t}$-direction get shorter, approaching zero coordinate-length at the singularity. Thus local light-cones in the interior of the EH are "infinitely wide" at the EH, and get "infinitely narrow" at the singularity, see Fig. 11.32.

We could say that the space-time G_{sb} is the worldview of an observer k suspended far away from the black hole. How far away? We measure "far" by multiples of the radius of the EH (which we chose to be 1 presently), and the farther away the suspended observer is, the more he will experience special relativistic space-time on his own worldline. He sees that inertial observers fall towards the black hole, but he never sees them reach the black hole, for him (k) they stay outside for the eternity of k. As they fall towards the black hole, they move slower and slower as they approach the event-horizon, and eventually they "freeze" onto the event-horizon. Also the wristwatches of the inertial observers tick slower and slower towards the EH, and "stop altogether at the EH". The same happens to the photons sent towards the EH in a radial direction: the photons slow up as they approach the EH, and eventually, and metaphorically, they "freeze" onto the event horizon. Our suspended observer k observes things this way both via photons (i.e. visually by his eyes) and via his coordinate system.

Since an in-falling inertial observer m lives for an infinite GFR-time, it is possible in theory that there is no upper bound for the time his wristwatch shows, i.e. that m also will experience that all his infinite time passes outside the EH. However, this is not the case: the wristwatch readings of an in-falling inertial observer m are bounded, e.g. the wristwatch of m may approach 12 as he falls in, tick slower and slower and never reach 12 o'clock, cf. Fig. 11.32. So

what happens in the in-falling inertial observer's own worldview or spaceship when his wristwatch reaches 12 o'clock?

We will see that he falls through the EH into the interior of the BH. For our suspended observer k, the interior of the BH is not visible by photons, he cannot get information about the inside of the BH while suspended. However, he may wonder what might be inside of that "big black ball", i.e. the EH. While suspended he cannot find out the answer. But, in principle, he may choose to fall in, and because of this, the interior of the BH is a "reality" for him.

In many ways this worldview G_{sb} of the suspended observer k is similar to the worldview of G_{ua} of an accelerated observer in special relativity. However, the following is an important and significant difference between G_{ua} and G_{sb}. Assume two inertial observers m, h fall radially into the black hole, in the same "path", i.e. in the same direction, and they started to fall at the same GFR-time, and close to each other. During their fall, one of them, say m, measures the distance between them by sending photons to h which it mirrors back to m. Then m measures constantly the time it takes for the photon to get mirrored back. The result of this measurement is called radar-distance. In G_{ua} this radar-distance remains the same, since the distance of parallel geodesics in G_{sr} does not change (this means that it is "flat"). However, in G_{sb}, m will find that the radar-distance between him and h is growing; and since G_{sb} is the relativistic version of a gravitational source, this is expected to be so because in Newtonian gravity, things closer to a gravitational source fall faster than things more distant. Technically speaking, timelike geodesics that started out in a parallel way, will increase their distance from each other in the tr-plane; the space-time G_{sb} is curved.[11] This shows that there is no partial isomorphism between G_{sb} and G_{sr} with an open domain.

Worldview of an observer falling into the black hole:

In the worldview of an observer falling into the black hole, the worldline of the in-falling observer m ought to be a straight line parallel with the time-axis. Instead of aiming for this, it is more convenient to "re-coordinatize" G_{sb} in such a way that the worldline of a radially ingoing photon will be a straight line of slope 1. There are no essential differences between such a worldview and a worldview where the worldline of m would be a straight line.[12] In fact, since the EH and the singularity in its middle are special "marked" places in this worldview, it is quite natural to make a worldview where these "do not move" in the GFR.

[11] Parallel geodesics diverge means negative curvature. Hence the tr-plane of G_{sb} is negatively curved. This is an instance of the so-called *tidal forces* which G_{sb} (and, in general, GR) inherited from the Newtonian theory of gravity.

[12] "parallel with the time-axis" is inessential here, since it is easy to rotate a picture.

In G_{sb}, the worldline of a radially in-falling photon in the tr-plane and outside the EH is $\{\langle -r - \ln(r-1) + \mathsf{constant}, r\rangle : r > 1\}$, and the worldline of a photon inside the EH is $\{\langle -r - \ln(1-r) + \mathsf{constant}, r\rangle : r > 1\}$. (This is not difficult to show by using the definition of a photonlike curve given in Sec. 3.3, and by knowing that the worldline is a subset of the tr-plane.) Thus the following simple (partial) function $Iso : \mathbb{R}^4 \to \mathbb{R}^4$ will take these photon worldlines to straight lines of slope 1:

$$Iso(t, x, y, z) := \langle t + \ln|r-1|, x, y, z\rangle \quad \text{where } r = \sqrt{x^2 + y^2 + z^2}.$$

Let us look for the isomorphic image of G_{sb} along Iso. By Theorem 11.24(ii) (p. 675), the simplest way of defining the isomorphic image $\mathsf{G}^0_{ef} = \langle M', L'\rangle$ of $\mathsf{G}_{sb} = \langle M, L\rangle$ is by letting $L'(Iso(p)) = \mathsf{D}(Iso)_p \circ L_p$ for all $p \in M$. By doing so we get the following definition (see Fig. 11.33).

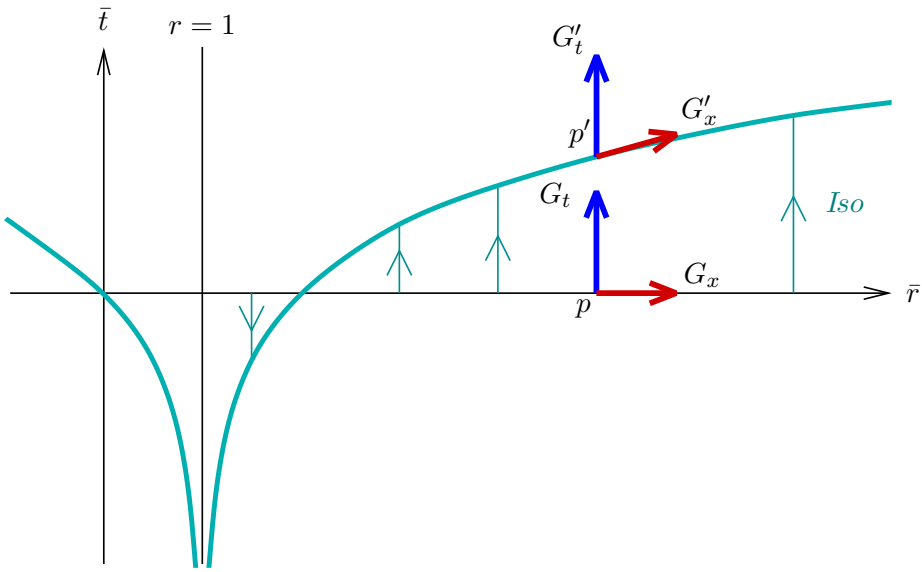


Figure 11.33. Eddington-Finkelstein re-coordinatization as an isomorphism between space-times.

The *Schwarzschild black hole in Eddington-Finkelstein coordinates* is $\mathsf{G}^0_{ef} = \langle M', L'\rangle$ where $M' = \{p \in \mathbb{R}^4 : r(p) \neq 0, 1\}$, and L' is specified by the vector-tetrad $G'(p) = \langle G'_t(p), G'_x(p), G'_y(p), G'_x(p)\rangle$ where

$$G'_x(p) = \sqrt{\tfrac{1}{r(r-1)}}\mathbf{1}_t + \sqrt{\tfrac{r-1}{r}}\mathbf{1}_r, \quad G'_t(p) = G_t(p) \text{ for } r > 1,$$

$$G'_t(p) = \sqrt{\tfrac{1}{r(1-r)}}\mathbf{1}_t - \sqrt{\tfrac{1-r}{r}}\mathbf{1}_r, \quad G'_x(p) = G_x(p) \text{ for } r < 1, \text{ and}$$

$$G'_y(p) = G_y(p), \quad G'_z(p) = G_z(p) \text{ for all } r.$$

(A few words on how we got $G'(p)$: Assume $r = r(p) > 1$ and let $p' = \mathsf{Iso}(p)$. Now $\mathsf{D}(Iso)_p$ takes $\mathbf{1}_t + p$ to $\mathbf{1}_t + p'$ and $\mathbf{1}_r + p$ to $\mathbf{1}_r + \frac{1}{r-1}\mathbf{1}_t + p'$. It can be seen in Fig. 11.33 that we get $G'_x(p)$ by considering its "slope (relative to $\mathbf{1}_x$)" and by considering the length of its "$\overline{r}$-projection". The slope is given by the derivative of *Iso*, thus it is $\frac{1}{r-1}$, and the $\overline{r}$-projection of $G'_x(p)$ is $G_x(p) = \frac{r-1}{r} * \mathbf{1}_r$. We obtain $\frac{1}{r-1}|G_x| = \frac{1}{r(r-1)}$ as the t-component of G'_x. The case $r < 1$ is analogous. G'_t, G'_y, G'_z are obtained similarly.) Thus, $G'(p)$ specifies the local LFR frame at p, see Fig. 11.34.

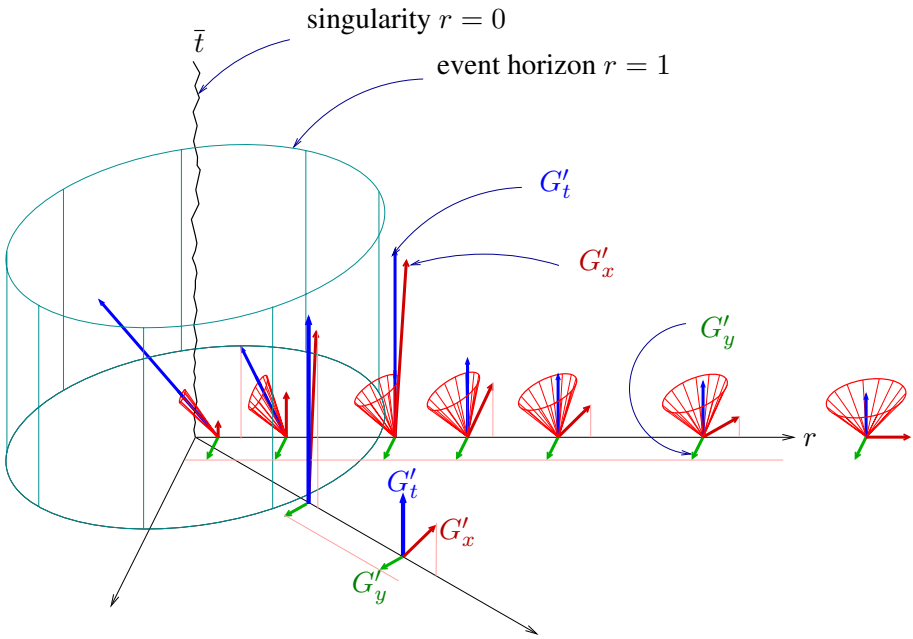


Figure 11.34. Schwarzschild space-time in Eddington-Finkelstein coordinates.

To see what the metric-tensor field $\overline{\mathsf{g}}_{ef}$ of G_{ef} is we have to "look at the coordinate unit vectors $\mathbf{1}_i$" with the eye of this LFR. This was explained in Sec. 3.4, on p. 672.

In the coordinate system specified by $G'(p)$, the coordinates of $\mathbf{1}_t$ and $\mathbf{1}_r$ are $\langle\sqrt{\frac{r-1}{r}}, 0, 0, 0\rangle$ and $\langle -\sqrt{\frac{1}{r(r-1)}}, \sqrt{\frac{r}{r-1}}, 0, 0\rangle$ respectively. Hence for $r = r(p) > 1$,

$$\mathsf{g}_{tt}(p) = \mathsf{g}_p(\mathbf{1}_t, \mathbf{1}_t) = (\sqrt{\tfrac{r-1}{r}})^2 = \tfrac{r-1}{r},$$

$$\mathsf{g}_{rr}(p) = \mathsf{g}_p(\mathbf{1}_r, \mathbf{1}_r) = (-\sqrt{\tfrac{1}{r(r-1)}})^2 - (\sqrt{\tfrac{r}{r-1}})^2 = -(1+\tfrac{1}{r}), \quad \text{and}$$

$$\mathsf{g}_{tr}(p) = \mathsf{g}_{rt}(p) = \mathsf{g}_p(\mathbf{1}_t, \mathbf{1}_r) = \sqrt{\tfrac{r-1}{r}} * -\sqrt{\tfrac{1}{r(r-1)}} = -\tfrac{1}{r}.$$

For $r < 1$ we obtain the same final values for $\mathrm{g}_{ij}(p)$. For this reason, the metric-tensor field $\overline{\mathsf{g}}_{ef}$ belonging to G^0_{ef} is given by the following line-element

$$\mathsf{d}s^2 = (1-\tfrac{1}{r})\mathsf{d}t^2 - \tfrac{2}{r}\mathsf{d}t\mathsf{d}r - (1+\tfrac{1}{r})\mathsf{d}r^2 + r^2\mathsf{d}\varphi^2.$$

In this metric-tensor, $\mathsf{g}_{tr} \neq 0$ because the coordinate unit vectors $\mathbf{1}_t, \mathbf{1}_r$ are not orthogonal in the eye of the local LFR specified by the vector-tetrad $G'(p)$, see Figs. 11.33, 11.34. This, $\mathsf{g}_{tr} \neq 0$, means that the light-cones in the tr-planes are tilted as illustrated in Figs. 11.34, 11.35.

We can see that the above $\overline{\mathsf{g}}_{ef}$ can smoothly be extended to the event horizon, i.e. to $EH = \{p \in \mathbb{R}^4 : r(p) = 1\}$. The reason for this is that the above formula for $\overline{\mathsf{g}}_{ef}$ is not degenerate for $r = 1$. Hence we can extend G^0_{ef} to EH, and this way we get *Eddington-Finkelstein space-time*, in short *EF-space-time*, $\mathsf{G}_{ef} = \langle M_{ef}, \overline{\mathsf{g}}_{ef}\rangle$ where $M_{ef} = \{p \in \mathbb{R}^4 : r(p) \neq 0\}$. This is an extension of G^0_{ef} and *Iso* is a partial isomorphism between G_{sb} and G_{ef}. The event horizon EH is part of the space-time here; and in fact the EH in a tr-plane is the worldline of a photon! This extended G_{ef} explains what happens on the event horizon and shows how the inside of the BH can be connected to the outside. (We note that the above given G' cannot be smoothly extended to EH, but one can smoothly change G' to G'' which gives the same metric-tensor field and which can be smoothly extended to EH.)

What will an in-falling inertial observer experience in G_{ef}? Throughout we assume that the BH in question is big enough so that the tidal forces on the EH and also well inside the EH are negligible. The present "animation" is based largely on G_{ef} in Fig. 11.35, but also Fig. 11.32, Fig. 11.36 are taken into account. So, it is useful to consult these figures with an emphasis on Fig. 11.35 before reading on. For visualizability, we assume that the BH is like the ones in centers of galaxies in that there are some stars (suns) orbiting our BH. So there is the EH, outside that there are the nearby suns orbiting the EH, and far away, there are the distant galaxies. As explained in Rindler, 2001, Sec. 12.5, pp. 267–271, Fig. 12.6, it is observationally possible to decorate the exterior of our BH with a latticework or "scaffolding" consisting of suspended observers (spaceships using fuel for maintaining their latitude) which surround the EH (only the exterior) and which maintain constant radar-distance from each other by using rockets. Gyroscopes are used to avoid rotation. We will think of these suspended observers as milestones, telling our in-falling observer where he is and what his speed is. It is impossible to have suspended observers on the EH or inside the EH, so one sign telling the in-falling observer that he is already inside will be the nonexistence or disappearance of the milestones.

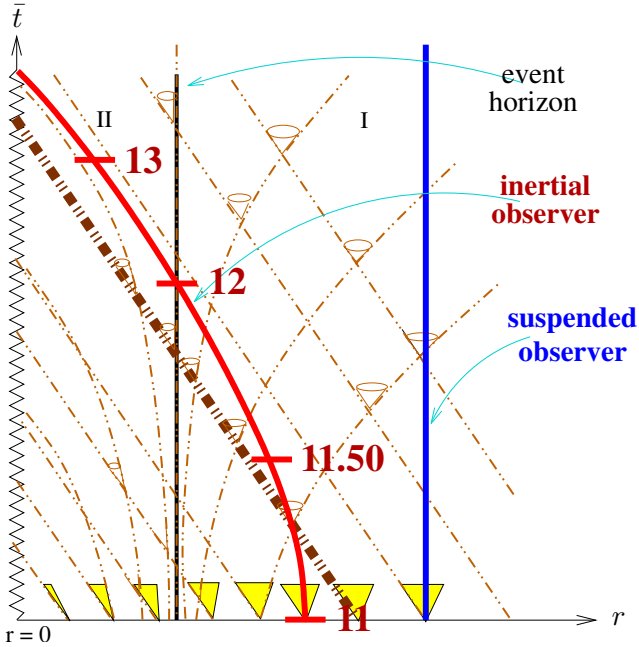


Figure 11.35. The "tr-slice" of space-time of Schwarzschild black hole in Eddington-Finkelstein coordinates.

As the in-falling observer m approaches the EH, he will see the milestones flashing by him, so to speak, faster and faster approaching the speed of light. If there were a milestone on the EH, then it would flash by (i.e. move relative to the observer) with the speed of light. However, this milestone cannot be realized by an observer. When all the milestones have flashed by our in-falling m, he will notice that there are no more milestones and even the BH has disappeared. Then m finds himself in basically empty flat-looking space[13] with no BH and no singularity in any of his spatial directions. More precisely, if he knows what to look for, then he can still observe some traces of the EH just as we can "see" our Big Bang (via the cosmic microwave background radiation) but it is all in the past, gone so to speak, and not influenceable causally. Like history is not changeable. The nearby suns (say, of different colors for fun) are also visible, even moving as m watches them, but they are like ghosts, their light comes from *before* the (Big Bang like) EH and they are causally not "touchable", since they all are in the distant past. All of the outside world, even the future

[13]Space may become flat (inside EH, of course), space-time remains curved.

of the suspended observers outside the EH, are in the past for m inside the BH, moving, living, dynamical, and changing but in the distant past before the Big Bang like EH, causally unreachable. What is interesting about this is that it is tempting to say that for m safely inside the EH the exterior of the EH does not exist. He is in a different universe, period. But that would not be the complete truth. Namely, the nearby suns circling the BH are still visible for m, they are just not influenceable causally. For more on this experience of seeing several universes via a BH we refer to the professional physics movie "Falling into a BH" (Hamilton, 2001), and the Prologue of Thorne, 1994. What we described so far is nothing but a decoding of the space-time diagrams Figs. 11.32-11.36 and of the metric of G_{ef}. This is the meaning of the mathematical expression of saying "space and time gets interchanged". Soon we will discuss more subtle BH's where m can avoid the fate of eventually hitting the singularity.

If we want to concentrate on the causal structure of a space-time, e.g. of the Eddington-Finkelstein black hole in Fig. 11.35, then that can be represented more compactly by a so-called conformal diagram (or Penrose diagram) of G_{ef}. Such a conformal diagram (of Fig. 11.35) is represented in Fig. 11.36. In a conformal diagram, photonlike geodesics are represented as straight lines of slope 1 and local time flows upwards.

G_{sb} and G_{ef} are two worldviews of the Schwarzschild black hole, connected by *Iso*. With an analogous way as we obtained G_{ef} we can obtain a re-coordinatization where the worldlines of the outgoing photons will be straight lines of slope 1. In this worldview, the interior of EH will behave like a so-called white hole: things can come out of the interior but cannot move inside. If we go on completing the worldlines of observers when we can we will also obtain a so-called hypothetical dual universe. All of these fit into one world-view called the Kruskal-Szekeres space-time, whose conformal structure is illustrated in Fig. 11.36.

4.3 Double black holes, wormholes

After an observer falls into a Schwarzschild black hole, he has only a finite time to live inside, and he must meet the singularity. There are many more friendly kinds of black holes, where he can live for an infinite time inside the black hole, he can avoid meeting a singularity, and he can even come out into an asymptotically flat region. (The expression "wormhole" refers to this last property.) We briefly describe here two such black holes, the electrically charged black hole and the rotating black hole.

Electrically charged black holes

This black hole is also called *Reissner-Nordström black hole* in the literature. Its line-element is

$$\mathsf{d}s^2 = (1 - \tfrac{1}{r} + \tfrac{e}{r^2})\mathsf{d}t^2 - (1 - \tfrac{1}{r} + \tfrac{e}{r^2})^{-1}\mathsf{d}r^2 - r^2\mathsf{d}\varphi^2$$

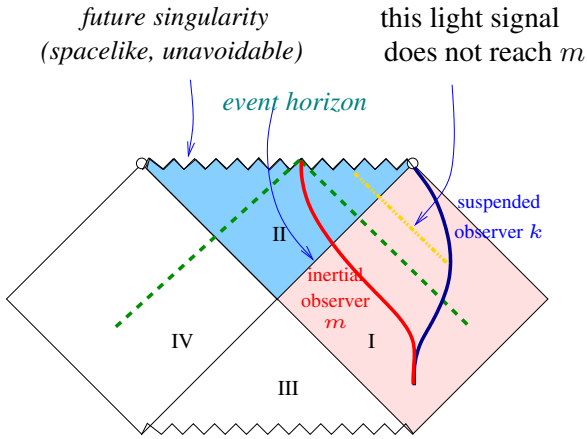


Figure 11.36. Conformal or Penrose diagram of "completed" Schwarzschild black hole. The shaded area consisting of blocks I, II is conformal diagram of EF-black hole. Region I is the exterior of the BH, region II is the interior of BH, region III is the white hole, and region IV is the dual universe. In some sense, regions III, IV may or may not exist but regions I, II have a stronger ontological status, they probably exist.

where $0 \leq e < \frac{1}{4}$. (Notice the strong analogy with Schwarzschild space-time G_{sb}.) Here, e represents the square of the electric charge. In this space-time $r = 0$ is the singularity, and there are two event horizons at

$$r^- = \tfrac{1}{2} - \sqrt{\tfrac{1}{4} - e} \qquad \text{and} \qquad r^+ = \tfrac{1}{2} + \sqrt{\tfrac{1}{4} - e}\ .$$

The exterior of the outer event horizon is similar to the Schwarzschild black hole: the light-cones get narrower towards the outer EH, and they are "infinitely narrow" at the EH. Inside the outer EH, the space-time remains similar to G_{sb} till about halfway towards the inner EH: time and space get interchanged and as we move inwards, local time gets faster and faster. But after a while, local time begins to slow down again, and local time "stops" at the inner event horizon, where time and space get interchanged once more. The innermost part, the interior of the inner event horizon, is similar somewhat to the exterior of the outer EH, but time runs faster and faster towards the singularity, beyond any limit. The singularity can be avoided in the inside of the black hole, the in-falling observer can "live forever".[14] The coordinatization represented by the above line-element is analogous to the Schwarzschild coordinatization of simple black hole, where the events of the in-falling inertial observers' entering the outer EH

[14]Actually, it is extremely difficult to go near the singularity (because of the repelling effect), so the in-falling observer is safe, will not be hurt by the BH.

are not included. An Eddington-Finkelstein-type re-coordinatization of the space-time where the worldlines of the ingoing photons are straight lines of slope 1 is illustrated in Fig. 11.37.

The conformal diagram of the electrically charged BH is shown in Fig. 11.39. An observer falling into this BH may come out to an asymptotically flat region (after crossing the EH's) as indicated in Fig. 11.39.

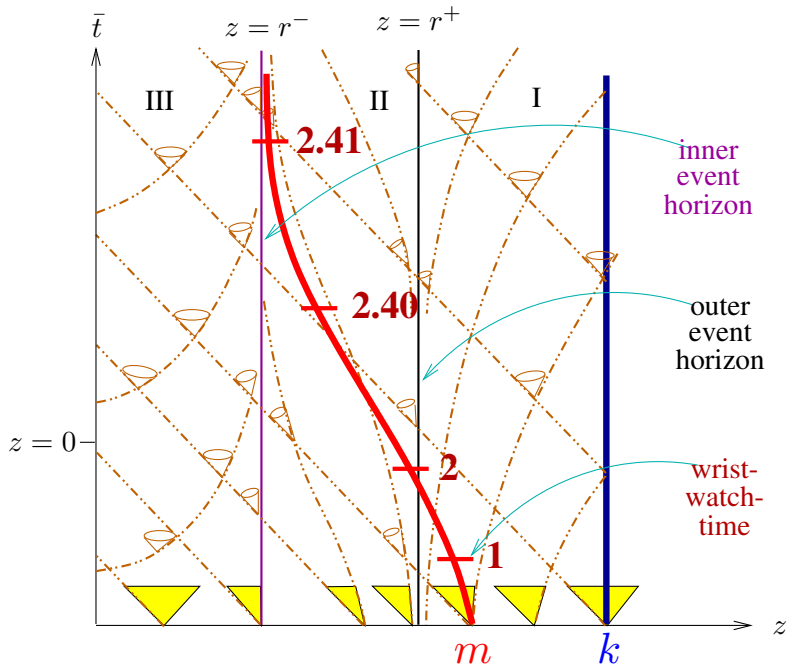


Figure 11.37. The "tr-slice" of electrically charged black hole. (Also the "tz-slice" of space-time of slowly rotating black hole in coordinates where z is the axis of rotation of black hole.) r^+ is the outer event horizon, r^- is the inner event horizon, $z = 0$ is the "center" of the black hole. The tilting of the light-cones indicates that not even light can escape through these horizons. That there is an outward push counteracting gravity can be seen via the shape of the light-cones in region III (central region of the black hole). The time measured by m is finite (measured between an event outside the inner EH and the event when m meets the inner event horizon) while the time measured by k is infinite.

Rotating (spinning) black holes

The space-times of slowly rotating black holes, called slow *Kerr space- times* in the literature, are similar to the electrically charged ones in that there are two EH's. We can think that the second, inner EH is the result of a repelling force overtaking the attraction of gravity. In the case of electrically charged black holes, the repelling force can be thought of, roughly, as the result of an excess of electrons (or protons) "in the BH", cf., e.g., d'Inverno, 1983, pp. 239–244 or Hawking and Ellis, 1973, p. 156 for a more careful explanation. In the case of rotating black holes, the repelling force can be thought of as the centrifugal force of rotation. The metric-tensor $\mathsf{g}(p)$ of Kerr black hole at $p = (t, r, \varphi, \vartheta)$ is given by the 4 by 4 matrix

$$\mathsf{g}_{Kerr} = \begin{pmatrix} -1+\mu & 0 & -\mu a \sin^2\vartheta & 0 \\ 0 & \Sigma/\Delta & 0 & 0 \\ -\mu a \sin^2\vartheta & 0 & \mathsf{g}_{\varphi\varphi} & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix},$$

where $\Sigma = r^2 + a^2\cos^2\vartheta$, $\Delta = r^2 - \mathsf{M}r + a^2$, $\mu = \mathsf{M}r/\Sigma$, and $\mathsf{g}_{\varphi\varphi} = (r^2 + a^2 + \mu a^2 \sin^2\vartheta)\sin^2\vartheta$. We used the so-called Boyer–Lindquist coordinates $(t, r, \varphi, \vartheta)$ where $(t, r, \varphi, \vartheta)$ are kind of polar-cylindric coordinates, r being radius (to be precise, r is the logarithm of the radius) and φ, ϑ being angular coordinates like η and θ were on p. 686. In the Kerr metric, g_{Kerr}, $\mathsf{M}, a \in \mathbb{R}$ are parameters, M corresponding to mass and a to the angular momentum of the rotating singularity. Indeed, choosing $a = 0$ yields the metric of the simple Schwarzschild BH. A two-dimensional slice of a slowly rotating black hole is very similar to the one in Fig. 11.37, and a "spatial" representation is in Fig. 11.38.

We meet two interesting features in these black holes. The first interesting feature is that there are so-called Malament-Hogarth events in these space-times. An event e is called a *Malament-Hogarth event* if in the causal past of e there is a time-faithful curve which is infinite in the future direction. The words "past" and "future" are important here, these refer to a time-orientation of the space-time, as follows. A *time-orientation* on a GR space-time $\langle M, L\rangle$ is a smooth vector-field each member of which is timelike (formally, a time-orientation is a smooth $T : M \to \mathbb{R}^4$ such that $\boldsymbol{\mu}_p(T(p) - p) > 0$ for all $p \in M$). All the GR space-times mentioned in Sec. 4 have natural time-orientations. Given a time-orientation, the notion of a future-oriented timelike curve can be defined. The causal past of an event e is defined to be the set of events e' which can be connected with e by a future-oriented timelike curve such that e' is "earlier" in this curve than e is.

One could think that in Malament-Hogarth events "actual infinity" is an observable physical reality. This phenomenon raises lots of intriguing questions to think over and has consequences even for the foundation of mathematics.

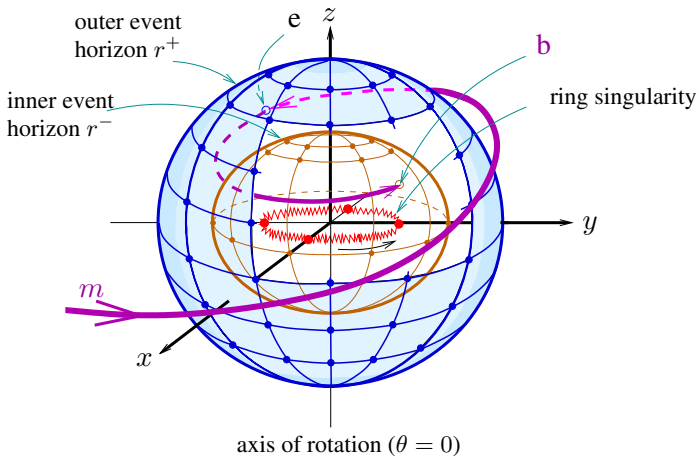


Figure 11.38. A slowly rotating (Kerr) black hole has two event horizons and a ring-shape singularity (the latter can be approximated/visualized as a ring of extremely dense and thin "wire"). The ring singularity is inside the inner event horizon in the "equatorial" plane of axes x, y. Time coordinate is suppressed. Fig. 11.37 is a space-time diagram of this with x, y suppressed. Rotation of ring is indicated by an arrow. Orbit of an in-falling observer m is indicated, it enters outer event horizon at point e, and meets inner event horizon at point b. For more on the basics of this figure cf. O'Neill, 1995, Fig. 2.2, p. 63.

For more on this we refer to Németi and Dávid, 2006, Németi and Andréka, 2006, and to Etesi and Németi, 2002. There are Malament-Hogarth events in both the charged and the rotating black holes, see Fig. 11.39.

Another intriguing feature is the presence of *closed timelike curves* (CTC's) in Kerr space-time. CTC's raise the question of time-travel into the past, and offer themselves for a logical treatment like the Liar Paradox. For more on this we refer to Earman, 1995. There are CTC's in the space-time of a rotating black hole, see e.g. O'Neill, 1995, pp. 76–77, Proposition 2.4.7, Wüthrich, 1999, Andréka et al., 2006c. There are many other kinds of space-times with CTC's, e.g. Tipler- van Stockum's rotating cylinder, Gödel's universe, the ones described in Thorne, 1994 and Novikov, 1998, to mention a few. Cf. Fig. 11.40.

4.4 Black holes with antigravity (i.e. with a cosmological constant Λ). Triple black holes

One can combine the idea of a BH with a universe in which the vacuum regions have a nonzero curvature characterized by Einstein's cosmological term $\Lambda \in \mathbb{R}$. Λ may be positive or negative, but $|\Lambda|$ is small. Recent cosmological observations suggest that Λ or something like it might be out there, i.e. might be important for understanding the acceleration of our expanding universe. The

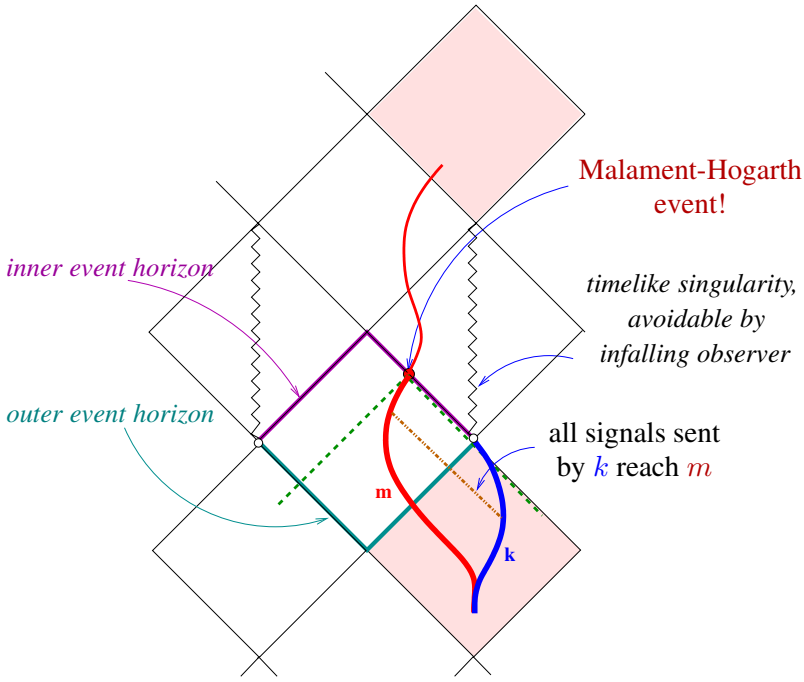


Figure 11.39. Penrose diagram of electrically charged black hole (and also of slowly rotating black hole). The red line represents a segment of the life-line of an in-falling inertial observer m, and the blue line represents the life-line of a suspended observer k. The time passed on the red line is finite, while the time passed on the blue line, i.e. for the suspended observer, is infinite. In principle, the in-falling observer has access in a finite wristwatch-time of his to all of the future history of the suspended observer k (an ultimate effect of "slow time" caused by BH's).

line-element is a generalization of the one of the charged BH (generalizing G_{sb}) on p. 697

$$\mathsf{d}s^2 = (1 - \tfrac{\mathsf{M}}{r} + \tfrac{e}{r^2} - \Lambda r^2)\mathsf{d}t^2 - (1 - \tfrac{\mathsf{M}}{r} + \tfrac{e}{r^2} - \Lambda r^2)^{-1}\mathsf{d}r^2 - r^2\mathsf{d}\varphi^2.$$

Here M is the mass of the BH, e is (square of) its electric charge, and Λ represents the hypothetical antigravitational property of intergalactic vacuum. The parameters M, e, Λ can be chosen independently of each other obtaining various kinds of special cases. $\Lambda > 0$ causes the timelike geodesics outside the outer EH behave as if an antigravitational force would be pushing them outwards, away from the EH, cf. Rindler, 2001, Sec. 14.4, pp. 304–306. If $\Lambda = 0$, we

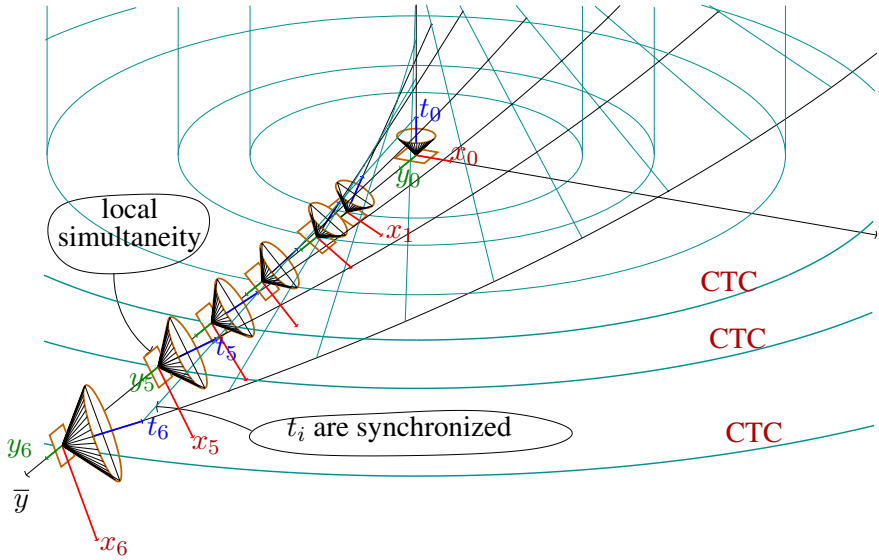


Figure 11.40. Starting point for Gödel's rotating cosmological model. This is a GR space-time, the vector fields and local light-cones representing the local special relativity frames (LFR's) are indicated. CTC's can be seen in the figure.

are in asymptotically flat universe, $\Lambda > 0$ means negative curvature, hence what is called de-Sitter universe, while $\Lambda < 0$ means positive curvature (for our vacuum), and so-called anti-de-Sitter universe.

The choice $\Lambda < 0$ causes distant clocks speed up[15] (via the $-\Lambda r^2$ term in g_{tt}), while small $\Lambda > 0$ causes them to run slow (assuming $\Lambda r^2 \leq 1$; at $\Lambda r^2 = 1$ there is a "coordinate-event horizon"[16] if $\mathsf{M} = e = 0$). The behavior (speed) of distant clocks determine the behavior of geodesics (gravitation) according to the same "logic" as explained at the Schwarzschild BH on pp. 689-690. The choice of $\mathsf{M} = e = 0$ yields *de-Sitter* and *anti-de-Sitter* space-times respectively, depending on the sign of Λ.

[15]Hence Malament-Hogarth computers breaking the Turing Barrier become possible, cf. Németi and Dávid, 2006, Hogarth, 2004.

[16]The event horizon at $r^2 = 1/\Lambda$ (i.e. where g_{tt} becomes 0 because of Λ) is in many respects like a huge Schwarzschild BH turned inside out, cf. Rindler, 2001, p. 306, lines 11–22. An electric BH in a de-Sitter space-time possesses three distinct event horizons, the innermost one caused, roughly, by e, the middle one caused, roughly, by M, and far out the outermost one cased by Λ. As we move away from the singuplarity lying on the time-axis $\overline{t}$, in the positive r direction, time and space get interchanged at crossing each one of the three event horizons.

4.5 Einstein's field equations

In the present work we concentrate on the space-time aspects of general relativity (GR). One of the reasons for this choice is that to study GR, it is reasonable to start with studying GR space-time, this enables one to study advanced and exotic examples of GR space-times like black holes, wormholes, cosmological models etc. as done e.g. in Taylor and Wheeler, 2000, and then turn to studying Einstein's field equations *EFE* and the rest of GR together with its "borderlines". This order is followed in, e.g., Penrose, 2004. About this possible continuation of studies we note the following. *EFE* is not a new axiom in the language of GR space-times restricting these. Instead, *EFE* is a definitional expansion of GR space-times in the sense of definability theory of mathematical logic (described in Sec. 2.6). *EFE* comes in two versions, a more flexible version, EFE^+, permitting the use of an extra parameter Λ for "fine-tuning" our space-time, and a less flexible one, *EFE*, in which $\Lambda = 0$ is assumed (or equivalently Λ is not used).

First we consider the $\Lambda = 0$ version. In this case, *EFE* is an explicit definition associating a tensor field denoted as $\langle T_{ij}(p) : p \in M\rangle$, or briefly T_{ij}, to every GR space-time $G = \langle M, \mathsf{g}\rangle$. From the logical point of view, T_{ij} is a brand new symbol not occurring in **Genrel** (or in the language of $\langle M, \mathsf{g}\rangle$ or $\langle M, L\rangle$). Hence **Genrel** + *EFE* can be regarded as a new theory expanding **Genrel** with new kinds of entities not mentioned in **Genrel** (or in its manifold oriented forms). Since *EFE* is an explicit definition (over **Genrel**) of the new entities called T_{ij}, the new theory **Genrel** + *EFE* is a conservative extension of **Genrel** (in the logical sense). This is the reason why we said earlier that *EFE* does not restrict the generality of **Genrel**, though it introduces a new linguistic (or conceptual) device to add such restrictions later if/when wanted and justified.

The physical role of *EFE* is the following. *EFE* helps us in elaborating the connections between **Genrel** and other physical theories (such as e.g. electrodynamics, or e.g. mechanics). This is so because the new concept T_{ij} (or new property T_{ij} of $\langle M, \mathsf{g}\rangle$) can be interpreted in the various physical theories as representing typical physical quantities like mass-energy-momentum density at points $p \in M$. In other words, the "new" tensor field T_{ij} can be regarded as associating various physical properties (or entities) to points p of the space-time under investigation. It is in this connection that T_{ij} makes it possible for related theories of physics to induce restrictions on the models of **Genrel** via **Genrel** + *EFE*.

The more flexible theory **Genrel** + *EFE*$^+$ is also a conservative extension of **Genrel** but in *EFE*$^+$ we introduce two new concepts, T_{ij} and Einstein's

cosmological parameter Λ . What is Λ? It intends to specify the curvature of vacuum.[17] What do we mean by referring to the vacuum, in **Genrel** there was no such concept as the vacuum. Again, using the concept of vacuum is connected to the physical interpretations of the theory. Roughly, vacuum consists of those points p of the space-time where $T_{ij}(p) = 0$. The assumption $\Lambda = 0$ amounts to assuming that the curvature of vacuum is the same as the curvature of Minkowski space-time, i.e. as that of special relativity. Intuitively, **Genrel** + *EFE*$^+$ permits us to choose an arbitrary but fixed value $\Lambda \in \mathbb{R}$ for the whole space-time. Usually $|\Lambda|$ is small. It was this extra flexibility which made it possible for Kurt Gödel to specify his rotating universe (Gödel, 1949) as a universe containing only pressureless dust.

EFE$^+$ can be written in the following form:

(*EFE*$^+$) $T_{ij} - \Lambda * \mathsf{g}_{ij} =$ expression of(g_{ij} and derivatives of g_{ij}).

Cf. e.g. Rindler, 2001, item 14.15, p. 303 or Wald, 1984, item 5.2.17, p. 99. In (*EFE*$^+$) we suppressed the constants deriving from units of measurement. By inspecting (*EFE*$^+$) above, one can see that instead of determining T_{ij} (matter-energy-momentum-etc density), it determines only the difference of T_{ij} and Λ, more precisely, it tells us the value of $T_{ij} - \Lambda\mathsf{g}_{ij}$ (from knowing g_{ij} and its behavior). Hence (*EFE*$^+$) leaves us a certain degree of freedom for distributing effects between T_{ij} and Λ. Further, g_{ij} occurs on both sides of the equation, hence (*EFE*$^+$) is only an implicit circumscription, not an explicit definition.

For completeness, we note that *EFE*$^+$ can also be used for a kind of classification of space-times, roughly, in terms of what they may "contain". An example is "vacuum space-times" which refer to space-times compatible with $T_{ij} = 0$ (uniformly). A complication here is that in principle the "division of labor" between T_{ij} and Λ is up to the interpreter's mind to choose. E.g. de-Sitter space-time (having a constant negative curvature) can be classified as a vacuum space-time with $\Lambda \neq 0$, or equivalently as one with $\Lambda = 0$ and T_{ij} nonzero. This classification can be further stretched to associate "realisticity" or "physicality" to space-times but such judgements often turn out to be subjective later. For illustration we note that Minkowski space-time is vacuum and so are G_{sb}, G_{ef}, the rotating BH's space-time, but the electrically charged BH's space-time is not vacuum (because the presence of electrical field at p implies $T_{ij}(p) \neq 0$).

5. Connections with the literature

Elaborating the logical foundations of relativity goes back to Hilbert's 6-th Problem. Most of the (logic oriented) work we are aware of concentrate

[17]The curvature of a GR space-time $G = \langle M, \mathsf{g}\rangle$ is a definable property of G. In more detail, the curvature tensor field of G is defined from the behavior of the geodesics of G.

on special relativity or on its fragments. Probably the first axiomatization for special relativity was given by Alfred Arthur Robb in 1914 (Robb, 1914), and his work is the model or starting point of many later axiomatizations. There are many works in which an axiom system for special relativity is given, a small sample (which is far from being complete) of these is: Robb, 1914, Reichenbach, 1969, Carathéodory, 1924, Alexandrov and his school starting with 1950 (Alexandrov, 1974, Guts, 1982), Suppes and his school starting with 1959 (Suppes, 1959, Suppes, 1968, Suppes, 1972), Szekeres, 1968, Winnie, 1977, Ax, 1978, Friedman, 1983, Mundy, 1986, Goldblatt, 1987, Schutz, 1997, Latzer, 1972. Of these, only Ax, 1978 and Goldblatt, 1987 are in first-order logic. These works usually stop with a kind of completeness theorem for their axiomatizations. What we call the analysis of the logical structure of relativity theory begins with proving such a completeness theorem but the real work comes afterwards, during which one often concludes that we have to change the axioms. Very roughly, one could phrase this as "we start off where the others stopped (namely, at completeness)". Most of this literature concentrate on what we call $\mathbf{Specrel}_0$, namely the causal fragment of special relativity without its metric aspect (which is present in **Specrel**). We note that there are interesting works connecting modal logic with special relativity, e.g. Goldblatt, 1980, van Benthem, 1983, p. 4,pp. 22-29, Casini, 2002, Shehtman and Shapirovsky, 2003. For more on the literature see Andréka et al., 2006b.

As a contrast with special relativity, we know only of a few attempts for providing a logical analysis of general relativity. Examples are Basri, 1966, Kronheimer and Penrose, 1967, Schröter, 2007, Busemann, 1967, Ehlers et al., 1972, Walker, 1959. None of these examples tries to stay within the framework of first-order logic (or even something like that, say, second-order logic) or attempts proving something like a completeness theorem. In Sec. 3.6 of the present work we propose a relatively simple first-order logic axiomatization **Genrel** for general relativistic space-times, and in Theorem 11.28 we formulate a completeness theorem for **Genrel**. What remains as a future research task is doing "reverse mathematics" for **Genrel**, i.e. providing a conceptual analysis for **Genrel** which would be analogous to the conceptual analysis provided for **Specrel** in Sec. 2 and in Andréka et al., 2002. Of course, a related future research task remains to push the present logic based conceptual analysis to the not yet existing theories conjectured to exist beyond general relativity like quantum gravity.

Acknowledgments

We are indebted to the participants of our 2004 and 2006 courses on "Relativity and Logic" at Eötvös University, Budapest, and especially to Zalán Gyenis, Ramón Horváth, Gergely Székely, and Renáta Tordai. We greatly profited from discussions and correspondence with Johan van Benthem, Gyula Dávid, Robin Hirsch, Péter Németi, Miklós Rédei, László E. Szabó and the participants of the

Oxford Space-time Workshop in 2004 and its follow-up conference in 2005. This research was supported by Hungarian Research grant OTKA T43242 as well as by a Bolyai Grant for Judit X. Madarász.

References

Aiello, M. and van Benthem, J. (2002). A modal walk through space. *Journal of Applied Non-Classical Logics*, 12(3–4):319–363.

Alexandrov, A. D. (1974). On foundations of space-time geometry. I. *Soviet Math. Dokl.*, 15:1543–1547.

Andréka, H., Madarász, J. X., and Németi, I. (1998–2002). On the logical structure of relativity theories. Technical report, Rényi Institute of Mathematics, Budapest. http://www.math-inst.hu/pub/algebraic-logic/Contents.html.

Andréka, H., Madarász, J. X., and Németi, I. (1999). Logical analysis of special relativity theory. In Gerbrandy, J., Marx, M., de Rijke, M., and Venema, Y., editors, *Essays dedicated to Johan van Benthem on the occassion of his 50th birthday*. Vossiuspers, Amsterdam University Press. CD-ROM, ISBN: 90 5629 104 1, http://www.illc.uva.nl/j50.

Andréka, H., Madarász, J. X., and Németi, I. (2004). Logical analysis of relativity theories. In Hendricks, V., Neuhaus, F., Pedersen, S. A., Scheffler, U., and Wansing, H., editors, *First-order Logic Revisited*, pages 7–36. Logos Verlag, Berlin.

Andréka, H., Madarász, J. X., and Németi, I. (2006a). Logical axiomatizations of space-time. Course material on the Internet, http://ftp.math-inst.hu/pub/algebraic-logic/kurzus-2006/kurzus-h-2006.htm.

Andréka, H., Madarász, J. X., and Németi, I. (2006b). Logical axiomatizations of space-time. Samples from the literature. In Prékopa, A. and Molnár, E., editors, *Non-Euclidean Geometries: János Bolyai Memorial Volume*, volume 581 of *Mathematics and Its Applications*, pages 155–185. Springer Verlag.

Andréka, H., Németi, I., and Sain, I. (1982). A complete logic for reasoning about programs via nonstandard model theory, Parts I-II. *Theoretical Computer Science*, 17:193–212, 259–278.

Andréka, H., Németi, I., and Sain, I. (2001). Algebraic Logic. In Gabbay, D. M. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 2, pages 133–247. Kluwer Academic Publishers, second edition. See also www.math-inst.hu/pub/algebraic-logic/handbook.pdf.

Andréka, H., Németi, I., and Wüthrich, C. (2006c). A twist in the geometry of rotating black holes: seeking the cause of acausality. Manuscript, Budapest and Berne.

Ax, J. (1978). The elementary foundations of space-time. *Found. Phys.*, 8(7–8):507–546.

Barbour, J. B. (1989). *Absolute or relative motion?* Cambridge University Press.

Basri, S. (1966). *A deductive theory of space and time*. North-Holland, Amsterdam.

Busemann, H. (1967). *Time-like spaces*, volume 53 of *Dissertationes Math. (Rozprawy Math.)*. Mathematical Istitute of Polish Academy of Sci.

Carathéodory, C. (1924). Zur Axiomatik der speziellen Relativitätstheorie. *Sitzungsber. phys. math.*, 14.:12–27.

Casini, H. (2002). The logic of causally closed space-time subsets. *Classical and Quantum Gravity*, 19(24):6389–6404. http://arxiv.org/abs/gr-qc/0205013.

Chang, C. C. and Keisler, H. J. (1973). *Model theory*. North-Holland.

d'Inverno, R. (1983). *Introducing Einstein's Relativity*. Oxford University Press.

Earman, J. (1995). *Bangs, crunches, whimpers, and shrieks. Singularities and acausalities in relativistic spacetimes*. Oxford University Press, Oxford.

Ehlers, J., Pirani, F. A. E., and Shild, A. (1972). The geometry of free fall and light propagation. In *General relativity, Papers in Honor of J.L. Synge*, pages 63–84. Clarendon Press, Oxford.

Einstein, A. (1961). *Relativity (The special and the general theory)*. Wings Books, New York, Avenel, New Jersey.

Etesi, G. and Németi, I. (2002). Non-Turing computations via Malament-Hogarth space-times. *International Journal of Theoretical Physics*, 41(2):341–370.

Friedman, H. (2004). On foundational thinking 1, Posting in FOM (Foundations of Mathematics). Archives www.cs.nyu.edu.

Friedman, M. (1983). *Foundations of Space-Time Theories. Relativistic Physics and Philosophy of Science*. Princeton University Press.

Gödel, K. (1949). An example of a new type of cosmological solutions of Einstein's field equations of gravitation. *Reviews of Modern Physics*, 21: 447–450.

Goldblatt, R. (1980). Diodorean modality in Minkowski spacetime. *Studia Logica*, 39:219–236.

Goldblatt, R. (1987). *Orthogonality and space-time Geometry*. Springer-Verlag.

Guts, A. K. (1982). Axiomatic relativity theory. *Russian Math. Survey*, 37(2): 41–89.

Hamilton, A. (1997-2001). Falling into a black hole. Internet page, http://casa.colorado.edu/ ajsh/schw.shtml.

Hawking, S. W. and Ellis, G. F. R. (1973). *The large scale structure of space-time*. Cambridge University Press.

Henkin, L., Monk, J. D., and Tarski, A. (1985). *Cylindric Algebras Parts I, II*. North-Holland, Amsterdam.

Henkin, L., Monk, J. D., Tarski, A., Andréka, H., and Németi, I. (1981). *Cylindric Set Algebras*, volume 883 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Hilbert, D. (1899/1977). *Grundlagen der Geometrie*. Lepzig / B. G. Teubner Verlag, Stuttgart.

Hirsch, R. and Hodkinson, I. (2002). *Relation algebras by games*. North-Holland.

Hodges, W. (1993). *Model theory*. Cambridge University Press.

Hogarth, M. L. (2004). Deciding arithmetic using SAD computers. *Brit. J. Phil. Sci.*, 55:681–691.

Horváth, R. (2005). An Alexandrov-Zeeman type theorem and relativity theory. Paper for scientific student contest, Eötvös Loránd University, Budapest.

Kronheimer, E. H. and Penrose, R. (1967). On the structure of causal spaces. *Proc. Camb. Phil. Soc.*, 63:481–501.

Latzer, R. W. (1972). Nondirected light signals and the structure of time. *Synthese*, 24:236–280.

Madarász, J. X. (2002). *Logic and Relativity (in the light of definability theory)*. PhD thesis, ELTE, Budapest. http://www.math-inst.hu/pub/algebraic-logic/Contents.html.

Madarász, J. X., Németi, I., and Székely, G. (2006a). First-order logic foundation of relativity theories. In *New Logics for the XXIst Century II, Mathematical Problems from Applied Logics*, volume 5 of *International Mathematical Series*. Springer. To appear. philsci-archive.pitt.edu/archive/00002726/.

Madarász, J. X., Németi, I., and Székely, G. (2006b). Twin paradox and the logical foundation of relativity theory. *Foundations of Physics*, 36(5):681–714. www.arxiv.org/abs/gr-qc/0504118.

Madarász, J. X., Németi, I., and Tőke, Cs. (2004). On generalizing the logic-approach to space-time towards general relativity: first steps. In Hendricks, V., Neuhaus, F., Pedersen, S. A., Scheffler, U., and Wansing, H., editors, *First-order Logic Revisited*, pages 225–268. Logos Verlag, Berlin.

Makkai, M. (1993). *Duality and definability in first order logic*. Number 503 in Memoirs of the AMS. AMS.

Misner, C. W., Thorne, K. S., and Wheeler, J. A. (1970). *Gravitation*. Freeman and Co, New York. Twentieth Printing 1997.

Mundy, J. (1986). The philosophical content of Minkowski geometry. *Britisch J. Philos. Sci.*, 37(1):25–54.

Németi, I. and Andréka, H. (2006). Can general relativistic computers break the Turing barrier? In Beckmann, A. Berger, U. Loewe, B. and Tucker, J. V. editors, *Logical Approaches to Computational Barriers, Second Conference on Computability in Europe, CiE 2006, Swansea, UK, July 2006, Proceedings*, volume 3988 of *Lecture Notes in Computer Science*, pages 398–412. Springer-Verlag, Berlin-Heidelberg.

Németi, I. and Dávid, Gy. (2006). Relativistic computers and the Turing barrier. *Applied Mathematics and Computation*, 178:118–142.

Nicholls, P., editor (1982). *The science in science fiction*. Crescent Books, New York.

Novikov, I. D. (1998). *The river of time*. Cambridge University Press.

O'Neill, B. (1995). *The geometry of Kerr black holes*. A K Peters.

Pambuccian, V. (2006). Alexandrov-Zeeman type theorems expressed in terms of definability. *Aequationes Mathematicae*. to appear.

Penrose, R. (2004). *The road to reality. A complete guide to the laws of the Universe*. Jonathan Cape, London.

Reichenbach, H. (1969). *Axiomatization of the theory of relativity*. University of California Press, Berkeley. Translated by M. Reichenbach. Original German edition published in 1924.

Rindler, W. (2001). *Relativity. Special, General and Cosmological*. Oxford University Press.

Robb, A. A. (1914). *A Theory of Time and Space*. Cambridge University Press. Revised edition, *Geometry of Time and Space*, published in 1936.

Sain, I. (1986). Nonstandard dynamic logic. Dissertation for candidate's degree, Hungarian Academy of Sciences, Budapest. In Hungarian.

Schröter, J. (2007). A new formulation of general relativity, Parts I-III. *Adv. Theor. Math. Phys.*, 1:1–70.

Schutz, J. W. (1997). *Independent axioms for Minkowski space-time*. Longoman, London.

Schwabhäuser, W., Szmielew, W., and Tarski, A. (1983). *Metamathematische Methoden in der Geometrie*. Springer-Verlag, Berlin. Hochschul text, viii+482pp.

Shehtman, V. and Shapirovsky, I. (2003). Chronological future modality in Minkowski space-time. In *Advances in Modal Logic-2002*, pages 437–459. King's College Publications, London.

Simpson, S. G., editor (2005). *Reverse Mathematics 2001*. Lecture Notes in Logic,. Association for Symbolic Logic. pp. x+401.

Smolin, L (2001). *Three roads to quantum gravity*. Basic Books.

Suppes, P. (1959). Axioms for relativistic kinematics with or without parity. In Henkin, L., Tarski, A., and Suppes, P., editors, *Symposium on the Axiomatic Method with Special Reference to Geometry and Physics*, pages 291–307. North-Holland.

Suppes, P. (1968). The desirability of formalization in science. *The Journal of Philosophy*, 27:651–664.

Suppes, P. (1972). Some open problems in the philosophy of space and time. *Synthese*, 24:298–316.

Szabó, L. E. (2002). *The Problem of Open Future, Determinism in the light of relativity and quantum theory*. Typotex, Budapest.

Szabó, L. E. (2006). Empiricist studies on special relativity theory. Book manuscript, Budapest.

Szczerba, L.W. (1970). Independence of Pasch's axiom. *Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys.*, 18:491–498.

Szekeres, G. (1968). Kinematic geometry: an axiomatic system for Minkowski space-time. *Journal of the Australian Mathematical Society*, 8:134–160.

Szmielew, W. (1974). The role of the Pasch axiom in the foundations of Euclidean Geometry. In *Proc. of the Tarski Symp. held in Berkeley in 1971*, pages 123–132. Providence, RI.

Tarski, A. (1959). What is elementary geometry? In Henkin, L., Tarski, A., and Suppes, P., editors, *Symposium on the Axiomatic Method with Special Reference to Geometry and Physics*, pages 16–29. North-Holland.

Tarski, A. and Givant, S. (1987). *A formalization of set theory without variables*, volume 41 of *AMS Colloquium Publications*. Providence, RI.

Taylor, E. F. and Wheeler, J. A. (2000). *Exploring Black Holes. Introduction to General Relativity*. Addison Wesley Longman.

Thorne, K. (1994). *Black holes and time warps. Einstein's outrageous legacy*. W. W. Norton and Company.

van Benthem, J. A. F. K. (1996). *Exploring logical dynamics*. Studies in Logic, Language and Information. CSLI Publications, Stanford.

van Benthem, J. F. A. K. (1982). The logical study of science. *Synthese*, 51: 431–472.

van Benthem, J. F. A. K. (1983). *The logic of time*, volume 156 of *Synthese Library*. Reidel Pub. Co., Dordrecht.

Wald, R. M. (1984). *General Relativity*. The University of Chicago Press.

Walker, A. G. (1959). Axioms for Cosmology. In Henkin, L., Tarski, A., and Suppes, P., editors, *Symposium on the Axiomatic Method with Special Reference to Geometry and Physics*, pages 308–321. North-Holland.

Winnie, J. A. (1977). The causal theory of space-time. In Earman, J. S., Glymour, C. N., and Stachel, J. J., editors, *Foundations of space-time Theories*, pages 134–205. University of Minnesota Press.

Wüthrich, C. (1999). On time machines in Kerr-Newman spacetime. Master's thesis, University of Berne.

Chapter 12

DISCRETE SPATIAL MODELS

Michael B. Smyth
Imperial College, London & University of Birmingham

Julian Webster
Imperial College, London

Second Reader

John G. Stell
University of Leeds

1. Introduction

By "discreteness" of a spatial model we generally understand that in any bounded neighbourhood, or (bounded) region, there are only finitely many elements of the carrier of the model. Thus the model should be (at least) locally finite. Another way of putting it is that the bounded regions should not be infinitely subdivisible.

As is well known, Aristotle argued (*Physics*: III,6; VI,2) that space and time intervals are potentially infinitely subdivisible. In 20th century mathematics, a similar position was adopted by Brouwer, in opposition to the actual infinities of Cantorian set theory. In proposing discrete spatial models, we embrace the possibility that space is (locally) neither actually not potentially infinite, but actually finite.

We mention three streams of motivation for adopting such a point of view. The first of these is found in the "tolerance space" tradition, going back to Poincaré, 1905. The idea here is that any perceptual (as opposed to mathematical or physical) "continuum" must be finite and, moreover, can be structured by means of a binary relation, with the aid of which some rudimentary topology can be developed. Poincaré proposed a definition (not ultimately successful) of

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 713–798.
© 2007 *Springer.*

dimension for such spaces. Many years later, Zeeman and his student Poston gave much fuller, technical accounts of the idea. More details, with references, are provided below.

Both Zeeman and Poston envisaged that the ideas might be applicable to, not only perceptual, but physical continua: see, for example, the concluding paragraphs of Zeeman, 1962. Around the same time (but independently and from an entirely different point of view), R. Penrose proposed certain discrete structures, mathematically not unlike tolerance spaces, as suitable building blocks for physical continua: "spin networks". In recent years spin networks have been taken up by several authors, along with other discrete structures such as "causal sets" (which are posets, ordered by a causality relation) as having the potential for quantizing space-time. According to the enthusiasts, the situation with regard to the discreteness or atomicity of space and time may be compared with that obtaining around a century ago, when physicists were at last beginning to accept the atomic theory of matter. Apart from Penrose, 1971 there do not, however, seem to be any canonical references in this area. A popular account which emphasizes the (claimed) discreteness of space and time is Smolin, 2001. As a sample of more technical references: Markopoulou and Smolin, 1997; Sorkin, 2002.

We do not attempt to deal with physics in the present work. It may be worth mentioning, though, that our theory of dimension (Sec. 6) is based to a considerable extent on that of Evako, who was motivated at least in part by physics. Also, we draw inspiration from quantum logic in developing our theory of regions (Sec. 5.7).

A third stream of work is more pertinent to our efforts here. This is the field known as *digital topology*. In digital image processing, the space and the images (or regions) have to be representable discretely. One may seek to develop a topology and geometry which can handle these discrete representations directly, without having to embed them in the traditional continua. The main aproach uses structures which are, in effect, tolerance spaces, in which the binary relation is (usually) adjacency of pixels rather than perceptual indistinguishability. A second approach (Khalimsky et al., 1990) uses true topological spaces rather than graphs/tolerance spaces. We regard these approaches as entirely compatible (see Sec. 2).

We describe now some of the distinctive features of the work which follows. In the "topological" part of the chapter (Sec. 2–6) the spaces are envisaged as belonging to the category of Čech closure spaces. Graphs and topological spaces alike are objects of this category. Formulating key definitions at the level of closure spaces means that we can have a uniform treatment for graphs (especially tolerance spaces) and topological spaces. An important example of this is the treatment of regular sets. The regular open and regular closed sets of topology are generalized to regular *interior* and *closure* sets at the level of

closure spaces. Besides closure spaces we consider (Sec. 4) the more familiar closure *systems* in which the operator, in contradistinction to the Čech closure, is idempotent but is not required to distribute over unions. This is done with a view to developing some techniques with which to study the collection of regular interior sets of a closure space. The important fact emerges that the regular sets form a complete ortholattice, but *not* in general a Boolean algebra (because not distributive). In contexts in which we wish to emphasize this ortholattice structure (and this is generally the case) we shall, as explained in Sec. 4, refer to the regular interior sets as the *orthoclosed* sets. In the main case of interest here, the tolerance space, some of the distinctions made in Sec. 4 are superfluous, and we have:

$$interior\ set \ = \ regular\ interior\ set \ = \ orthoclosed\ set.$$

These sets are intended to serve as the *regions* of our spaces.

The question immediately arises as to whether we can hope to work with a region lattice that is not distributive. The three extended examples in Sec. 5 are intended to assist in answering this question. The first of these examples deals with the "logic" in which ortholattices which are not distributive feature most prominently, namely quantum logic. In the quantum setting it turns out that the lattices in question are actually *orthomodular*, a condition which can be regarded as a weak form of distributivity. In Theorem 12.29 and Proposition 12.30 we have tried to summarize the results from the quantum structures literature that may have a bearing on our endeavours here. In the next two examples of section 5 we consider region connection theory, and especially some attempts which have been made to allow for discrete models. (In comparing this kind of work with tolerance spaces, the graph relation is now construed as connection.) The problem is that in the well known formulations of connection theory, whether by Whitehead, Clarke, or the authors of RCC, non-discreteness (infinite divisibility) is built in. Thus we are particularly interested in attempts by J. Stell and associates to develop a discrete connection theory. Two kinds of lattices (dual p-algebras and bi-Heyting algebras) are proposed by these authors as capable of supporting a discrete region theory. The bi-Heyting algebra in particular is claimed (as suggested originally by Lawvere) to be well suited for geometry as it permits a nice notion of *boundaries* of regions to be introduced. But there is a problem with this: no matter what the region, its boundary is always 0-dimensional. Thus the proposal seems to be ruled out on dimensional grounds.

In Sec. 6 we develop our theory of boundary and dimension. It builds on work of Evako and associates. In the simple case of tolerance spaces, the Evako dimension is in effect given by the size of the largest clique. But this means that products do not work well: in effect the dimension (strictly speaking, rank) of a product space is the product of their dimensions rather than the sum. Our

solution involves the idea that some cliques are better than others. (Not all cliques can be considered as true "cells" of the space.) Restricting attention to the "good" cliques, we get a poset whose length may be taken as the dimension of the space.

In Sec. 7 we begin to go beyond topology, specifically to convex and affine structure. It is now assumed that the elements of the tolerance space are themselves cells, rather than structureless points. Formally, a cell is just a finite subset of the ground set. Informally, we can think of a cell A as the relative interior of the convex hull of A (and the connection, or tolerance, relation as overlapping of these open cells). Two features of the material in Sec. 7 are particularly significant. It turns out that, when the space is realizable (in terms of finite subsets of a Euclidean space), the lattice of regions is always orthomodular—and thus has "plenty of distributivity" within it, even though the lattice as a whole is not distributive. An unsolved problem at present is whether this orthomodularity property holds for "cell spaces" in general, whether realizable or not. The significance of orthomodularity is illustrated by showing that it leads to a particularly simple definition of *triangulation* (Sec. 7.4).

The second feature of the cell space material to be highlighted here is the identification, given a few plausible axioms, of such spaces with oriented matroids. We begin to explain this identification in Sec. 7.3. The proof of the equivalence is then given in Sec. 9. This proof goes via a new axiomatization of oriented matroids (via "surrounding sets"); thus the main burden of the proof is to show that this axiomatization is actually equivalent to the standard ones. An important feature of this equivalencing of geometries is that it works with spherical versions of the spaces involved.

Along with sphericity comes a commitment to spaces that are finite, rather than just locally finite. (In topological terms we have: compact + locally finite $\Rightarrow$ finite.) This is fully apparent in the remainder of the chapter, in which we discuss what is by far the most developed approach to a purely combinatorial account of affine and convex structure: matroid theory.

Sec. 8 is an introduction to matroids and Sec. 9 an introduction to oriented matroids. Oriented matroids may be considered as combinatorial Euclidean geometries. Any finite subset of $\mathbb{R}^n$ inherits an oriented matroid structure, which captures its subspace geometry. Further, any oriented matroid—a purely combinatorial structure defined axiomatically with no reference to $\mathbb{R}^n$—can be realized as a collection of pseudo-spheres on the surface of an n-sphere. The theory illustrates, subject to a reservation to be mentioned shortly, the basic approach to space taken in this chapter.

To have a categorical axiomatization—that is, an axiom system with only one model—is not at all a desideratum in our approach. Rather what we have is that there are, at least potentially, infinitely many oriented matroids, some of which contain more structure than others. Category theory provides the language

in which to discuss this issue more adequately. The class of all finite spaces of a given type should form a category, with refinement (i.e. more structure) defined in terms of morphisms. Completeness of this category with respect to the appropriate (most usually, inverse) limits would be attained by adding in the infinite, classical models of the axioms. That such completeness would obtain is what we have, in a special case, called the "Correspondence Principle" (Sec. 2). It is perhaps worth mentioning that such a category of spaces is being proposed for its theoretical significance, rather than directly as a foundation for spatial computation. Programs are thought of as being written in terms of the spatial models themselves, the *objects* of the category. Although practical computational issues are not a major concern in this chapter, we illustrate the point with a short discussion of Knuth's work on oriented matroids (Sec. 10).

Now we have to admit that "the" category of spaces is (known and) available only in certain cases. For topology, such categories and constructions are certainly available; see Sec. 1 for references to our previous work in this area. For geometry, particularly as formulated in terms of oriented matroids, the main stumbling block at present is an almost complete lack of morphisms. (This is the "reservation" mentioned earlier.) In Sec. 10 we suggest that the way forward might be to now try to go beyond synthetic geometry and develop a finite algebraic theory of space. It seems anyway a natural progression from topology to geometry to algebra in seeing how far the combinatorial method of spatial representation can be pushed. The models in this theory are very likely to be *spherical*, which is why in Sec. 9 spherical oriented matroids are discussed more than flat ones. Basic to combinatorial sphericity is the notion of an *involution* (the involute of a point on the surface of a sphere is its antipode), and the natural geometric interpretation of a set together with an involution is a crosspolytope (rather than a simplex, which is the natural geometric interpretation of a set). The vertex-edge graph of the n-crosspolytope is the complete n-partite graph discussed in Sec. 1.

One might expect that symmetry is fundamental in any spherical formulation of space, and Sec. 10 concludes with an outline of recent work by Gelfand and others on *Coxeter matroids*, which is a generalization of matroids entirely in terms of Coxeter groups.

2. Preliminaries; correspondence principle

Following some basic definitions concerning graphs, this section introduces what we term the Correspondence Principle. This is essentially the idea that graph theory provides discrete counterparts for key topological notions. As a sample of the evidence for the validity of this principle, we present the graph-theoretic counterparts of two significant topological/metric space notions, namely the contractible space and the hyperconvex metric space. These

counterparts are: the dismantlable graph and the Helly graph. Definitions and examples of these counterparts are provided. We believe that both concepts are of importance in discrete spatial modelling, although for reasons of space we shall not develop this theme in this chapter. The class of dismantlable graphs will make a brief appearance in the discussion of surfaces in Sec. 6. The Helly concept as such will not be needed later on; however, the graph which serves us as the discrete counterpart of the n-cube, and which has a substantial role in what follows, is actually a characteristic Helly graph. (For the sense in which the discrete cube characterizes the class of Helly graphs, see Observation 2 preceding the statement of the Correspondence Principle below.)

A (di)graph (G, R) is for us simply a binary relation R on the set G (of *vertices*). Thus, our graphs do not admit multiple edges. A loop is present at a vertex x, or not, according to whether xRx or not. (We do not employ the conventional term "simple graph", as it blurs the distinction between reflexive and irreflexive relations.) We almost always want the relation R to be symmetric, and this is generally intended by the use of the term *graph*. Moreover we are principally interested in two cases:

1 The relation is reflexive. Use of the symbol $\sim$ for the relation indicates this case.

2 The relation is *ir*reflexive (no vertex is related to itself). The symbol $\perp$ will indicate this case.

Given a graph $(G, \sim)$, we sometimes use the notation *co*-G for the irreflexive graph with vertex set G and relation the complement of $\sim$; likewise if we start with a graph $(G, \perp)$.

A *(graph) morphism* is a relation-preserving map (on vertices). The product of graphs is for us the category product, i.e. $(G, R) \times (G', R')$ is $(G \times G', R \times R')$, where $(x, x')R \times R'(y, y')$ holds iff xRy and $x'R'y'$. (Many other notions of product are used by graph theorists.) We denote the m-path, that is, the graph with $m + 1$ vertices $v_0, v_1, \ldots, v_m$ and with $v_i R v_j$ iff $|i - j| \leq 1$, by I_m. Our reason for departing from the usual notation here is that the paths serve for us as "discrete (unit) intervals". In particular, we consider I_m^n (that is, the n-fold product of the path I_m) as a *digital n-cube*.

A second useful operation on graphs G, G' (whose vertex sets are assumed disjoint), sometimes called the *sum*, is defined as follows: $G * G'$ has vertex set $G \cup G'$. Every vertex of G is taken as adjacent to every vertex of G'. Additionally, two vertices belonging to the same summand (G, or G'), are adjacent in $G * G'$ exactly if they are adjacent in that summand.

We sometimes let the natural number n denote the (reflexive) graph with n vertices and no edges other than the self-loops. With this notation, the complete k-partite graph with k components, having $n_1, \ldots, n_k$ vertices respectively, can be expressed as $n_1 * \ldots * n_k$. The notation $K(n_1, \ldots, n_k)$, or some variant

thereof, is commonly used for this graph. The case where each $n_i = 2$ will be particularly important later.

A *clique* of a graph (G, R) is a set C of vertices such that, for every pair of distinct elements x, y of C, xRy. A *simplicial complex* is a collection Σ of finite subsets of a set S, including each singleton set, such that

$$A \subseteq B, B \in \Sigma \;\; \Rightarrow \;\; A \in \Sigma.$$

The cliques of any (locally finite) graph provide an example, indeed the main example that we are interested in. The simplicial complex Σ (over S) will be called *graph-like* if the simplexes of Σ coincide with the cliques of some graph (S, R). Here we can specify R to be the relation given by:

$$xRy \;\; \Leftrightarrow \;\; \exists A \in \Sigma.\, x \in A \;\wedge\; y \in A.$$

For any reflexive graph G we have the associated metric d_G, where $d_G(x, y)$ is the length of the shortest path (measured by the number of edges) from x to y. For any vertex $x \in G$ the *k-ball* $B(x, k)$ is the set $\{y|\; d_G(x, y) \leq k\}$.

In any digraph (G, R) we take the *neighbourhood* $N(v)$ of a vertex to be the set $\{x|vRx\}$. If G is a reflexive graph, this is the 1-ball $B(v, 1)$. Sometimes the *punctured* neighbourhood $N_0(v) = \{x|vRx\} \setminus \{v\}$ is required.

If G' is a subset of the set of vertices of the digraph (G, R), the *induced subgraph* on G' is simply the restriction (G', R') of G to G': that is, $vR'w$ in G' iff vRw in G. If G' is an induced subgraph of G, and there is in addition a surjective morphism from G onto G', we say that G' is a *retract* of G (and that the surjective morphism is a *retraction* from G to G'). Simple examples are provided by the wheels and cycles to be mentioned in a moment.

In keeping with our view of graphs as spatial models, we shall sometimes take the liberty of referring to reflexive graphs as "tolerance spaces", and to irreflexive graphs as "orthogonality spaces". There is in fact ample precedent for this kind of terminology. Poincaré, 1905 suggested that the "perceptual continuum" is, or should be, structured by means of a binary relation of *indiscernibility*, which is reflexive and symmetric, but not in general transitive. Having lain fallow for many years, the idea (along with the terminology of "tolerance spaces") was reintroduced by Zeeman around 1960: Zeeman, 1962. Under the name "fuzzy geometry" it was comprehensively developed, as a kind of alternative topology, by Poston in his thesis (Poston, 1971). For a concise treatment of tolerance spaces, and of their category **Tol** (the morphisms being the usual graph morphisms) see Sossinsky, 1986. Our own contributions to it are found in Smyth, 1995; Smyth, 1997. Another recent contribution is Georgatos, 2003.

The name "orthogonality space" for the irreflexive graph, on the other hand, has a precedent in discussions of quantum logic (a topic which we consider briefly in Sec. 5.1 below).

We illustrate the topological and metric aspects of graphs by an informal discussion of two (related) concepts. Almost nothing in the remaining sections depends on these examples, and we shall make use of some (standard) general terminology without troubling to define it.

The first concept is that of a *contractible* (also *dismantlable*) graph. The definition can be given in a way that reads exactly as in topology: A graph G is contractible if the identity function on (the vertices of) G is homotopic to a constant function. Several remarks are in order:

1 Homotopy for graphs can be treated in an entirely discrete fashion. Instead of the continuum $[0, 1]$ one may use the discrete intervals (paths) I_n: see for example Poston, 1971.

2 For finite reflexive graphs, an elementary definition, in recursive style, is available: G is *dismantlable* if there exists a vertex v of G such that, for some vertex u distinct from v, $N(v) \subseteq N(u)$, and $G - \{v\}$ is dismantlable. (A prime example of a "vertex dismantling" definition.)

3 Dismantlable *non*-reflexive graphs have recently been applied in modelling physical systems exhibiting "hard constraints": see Brightwell, 2000.

Simple examples of dismantlable and non-dismantlable graphs are provided by the *wheels* W_n and the *cycles* C_n ($n \geq 3$). Notice that $W_n = 1 * C_n$. The wheels and C_3 are dismantlable; the remaining cycles are non-dismantlable. As an illustration of previously defined terms, we also observe: for $n \geq 4$, C_n is an induced subgraph but not a retract of W_n, while I_2 is a retract of every such wheel and cycle.

(a) C_4 (b) W_5

Figure 12.1.

If we define a *cone* to be any (reflexive) graph of the form $1 * G$, then we evidently have that every cone is dismantlable. In fact, it is easy to see that if a

graph H is expressible as a sum $G * G'$ and at least one of the summands G, G' is dismantlable, then H is dismantlable.

The second concept has a metric aspect. A graph G is said to be a *Helly graph* if the set of balls of G has the Helly property: every collection of pairwise intersecting balls has a non-empty intersection. Some observations follow:

1 Every (finite) Helly graph is dismantlable: Quilliot, 1983; Bandelt and Pesch, 1989.

2 Helly graphs derive much of their significance from the following characterization: they are the injective objects in the category **Tol** (with respect to isometric embeddings). A more concrete characterizaton is: a graph is Helly if and only if it is a retract of a product of paths.

3 Helly graphs have their counterpart in classical analysis: the hyperconvex metric spaces. The details of this correspondence, together with a common generalization of the two concepts (namely, the hyperconvex *semi*-metric space), have been set out in Smyth and Tsaur, 2002.

We have been led by these, and many other, considerations to propose the following:

> CORRESPONDENCE PRINCIPLE: The central ideas of topology have discrete counterparts in graph theory. Moreover, the former arise by a limiting process from the latter.

"Ideas" here is meant to cover both definitions and theorems. For the second part of the Principle to make sense, we have to think of graphs and topological spaces as existing in the same category. There are several interesting possible choices for such a category. The best-known of these will make an appearance in the next section.

We should emphasize that, in the present chapter, we shall not be studying the limiting processes (especially inverse limits) needed to fully justify the Correspondence Principle. We have written extensively about such matters in previous work (Webster, 1997; Smyth and Webster, 2002; Smyth, 1995; Smyth, 1997). Here we shall illustrate, rather, the first part of the C.P. with a very simple example. The idea of the fixed point(s) of a function is certainly central in topology. What is the discrete counterpart? It turns out that we need to work with "almost fixed points" (or fixed points up to tolerance):

DEFINITION 12.1 *Let H be a graph. An* almost fixed point *of an endomorphism $h : H \to H$ is a vertex p such that $h(p) \sim p$. H is said to have the* almost fixed point property *(AFPP) if every endomorphism of H has an almost fixed point.*

EXAMPLE 12.2 *The 4-cycle does not have the AFPP (because of the mapping which sends each vertex to the diagonally opposite vertex), while the wheel has the AFPP.*

That the wheel has the AFPP is a special case of the following result:

THEOREM 12.3 *Every dismantlable (reflexive) graph has the AFPP.*

This result has been rediscovered a number of times. The first statement and proof of it seems to be Pultr, 1963. Independently, Poston, 1971 provided an elaborate treatment in terms of algebraic topology. A special case was considered by Rosenfeld, 1986. Generalizations involving multifunctions were studied in Tsaur and Smyth, 2001.

Notice that the AFPP is not hereditary (that is, it is not inherited by induced subgraphs). To see this, we need only look at the wheel and the cycle. Retracts, as well as products, are much better at preserving "topological" properties than are arbitrary induced subgraphs. In particular we have:

PROPOSITION 12.4

1 *Suppose that H is a retract of G by $r : G \to H$, $e : H \to G$, and that G has the AFPP. Then H has the AFPP.*

2 *If G_1, G_2 have the AFPP, then so does $G_1 \times G_2$.*

Proof

1 Assume that G has the AFPP. Suppose $h : H \to H$. Choose an almost fixed point p of the map $e \circ h \circ r : G \to G$. Then, since $p \sim e \circ h \circ r(p)$, we have $r(p) \sim r(e \circ h \circ r(p)) = (r \circ e)(h \circ r(p))$. Thus $r(p)$ is an almost fixed point of h.

2 Left to the reader.

QED

A class of graphs closed under products and retracts is known as a variety of graphs. Thus the graphs having the AFPP constitute a variety. It is also true that the classes of dismantlable and of Helly graphs are varieties.

More broadly, the category-theoretic aspect of graphs is of great importance, though neglected in most hitherto existing graph theory. This situation is beginning to be remedied, with the appearance of works such as Hell and Nešetřil, 2004. In terms of our view of graphs as discrete counterparts of topological spaces, graph morphisms correspond to continuous functions. (A precise formulation of this will be given in the following section.) Our brief treatment of dismantlable and Helly graphs in this section was intended to be illustrative of the significance of morphisms, as well as of the topological aspect of graphs.

The lack of any systematic treatment of morphisms is typical of the entire field (of discrete spatial models). Unfortunately we shall in this chapter be able to do little to remedy this situation. There is as yet, for example, no satisfactory category of oriented matroids. Questions of this kind constitute a substantial prospective area of research.

3. Čech closure spaces

As a convenient general setting within which we can study and compare spatial models of various kinds, whether discrete or continuous, we may take the formalism of *closure spaces*. A closure space is a set S equipped with an operator (the closure, Cl) acting on subsets of S. Closure spaces come in two main flavours. Precise details will be provided shortly (in this section and the following one) but a preliminary indication may be helpful. In the first flavour, we can think of $\mathrm{Cl}(S)$ as the set of elements *close* to the set S, in a suitable sense. The characteristic axiom here is:

$$\mathrm{Cl}(A \cup B) = \mathrm{Cl}(A) \cup \mathrm{Cl}(B).$$

(An element is close to $A \cup B$ if and only if it is close to A or close to B.) In the second flavour we typically have that $\mathrm{Cl}(S)$ is the set of elements directly or indirectly *dependent* on S. Here we invariably have idempotency:

$$\mathrm{Cl}(\mathrm{Cl}(S)) = \mathrm{Cl}(S),$$

together with (often) some more specific axiom reflecting the type of dependency involved.

The less familiar "distributive" (but non-idempotent) closure spaces are discussed in this section. The idempotent spaces appear in the next section. To avoid confusion, we shall refer to the idempotent spaces as closure *systems*.

Čech, 1966, demonstrated in great detail that much of general topology goes through without the assumption that the operation of closure is idempotent. It seems that he was motivated to do this by some examples in functional analysis. But it turns out that this generalization is exactly what we need if we seek to extend topology to "digital" spaces.

DEFINITION 12.5 *A (Čech)* closure space, (X, Cl), *is a set* X *equipped with a set operator* Cl *satisfying:*

1 $A \subseteq \mathrm{Cl}(A)$,

2 $\mathrm{Cl}(\emptyset) = \emptyset$,

3 $\mathrm{Cl}(A \cup B) = \mathrm{Cl}(A) \cup \mathrm{Cl}(B)$.

Closely associated with this notion of closure, we have those of *interior* and *neighbourhood*. Just as in ordinary topology, these three are interdefinable (at

least if we allow classical logic). Thus, in a closure space (X, Cl), we define the *interior* operation by:

$$\mathrm{Int}(A) = X - \mathrm{Cl}(X - A);$$

and we say that the subset B is a *neighbourhood* of A if

$$A \subseteq \mathrm{Int}(B).$$

For the *specialization* preorder $\leq_X$ of a closure space X we have:

$$x \leq_X y \;\; \Leftrightarrow \;\; (\forall B \subseteq X.\, x \in \mathrm{Int}(B) \Rightarrow y \in \mathrm{Int}(B)).$$

In words: every neighbourhood of x is a neighbourhood of y. Equivalently (check!) we can say:

$$x \leq_X y \;\; \Leftrightarrow \;\; \mathrm{Cl}(x) \subseteq \mathrm{Cl}(y);$$

or even more simply: $x \in \mathrm{Cl}(y)$. Given this, we naturally express T_0-separation by :

$$X \text{ is } T_0 \;\; \Leftrightarrow \;\; (x \equiv_X y \;\Rightarrow\; x = y).$$

("Identity of Indiscernibles.") Note that with this definition of T_0-separation we depart significantly from Čech, 1966. The details, together with our reasons for changing Čech's definition, are given in Smyth, 1995.

Continuity of functions is characterized via closure, interior, or neighbourhoods just as usual. (However, the definition via open sets or closed sets is not appropriate here.) Thus: the function $f : X \to Y$ is *continuous* if, for any $A \subseteq X$, $f(\mathrm{Cl}(A)) \subseteq \mathrm{Cl}(f(A))$.

EXAMPLE 12.6 *Any reflexive digraph (G, R) is a closure space, on taking the closure of a set of vertices to be given by:*

$$\mathrm{Cl}(A) = \{x \mid \exists y \in A.\, xRy\}.$$

The interior of a set B is then given by

$$\mathrm{Int}(B) = \{x \mid \forall y.\, xRy \Rightarrow y \in B\}.$$

A neighbourhood of a vertex v is any set of vertices which contains all the R-successors of v. It is easy to check that a mapping of digraphs is a graph morphism (relation-preserving map) if and only if it is continuous in the sense of closure spaces.

REMARK 12.7 *In case R is transitive (that is, the digraph is a pre-order), the closure is idempotent, and we have the usual* Alexandroff *topology of the pre-ordered set: that is, the topology obtained by taking the upper sets in the pre-order as open (equivalently, by taking the lower sets as closed).*

REMARK 12.8 *It can easily be shown that a closure space* (X, Cl) *is a graph (more precisely, is the closure space derived from a digraph as in the preceding example) if and only if each point of* X *has a smallest neighbourhood.*

In thus describing digraphs as closure spaces, we had to require reflexivity of the graphs on account of Axiom 1 of closure spaces. Much of the theory of closure spaces can still be carried through if we drop Axiom 1 (so that arbitrary digraphs are captured). We shall generally assume our (di)graphs to be reflexive, although this is in many cases unnecessary from the mathematical point of view.

4. Closure systems

4.1 Definitions; first examples

DEFINITION 12.9 *A* closure system (X, K) *is a set* X *together with a set operator* K *satisfying:*

1 $A \subseteq K(A)$,

2 $A \subseteq B \Rightarrow K(A) \subseteq K(B)$,

3 $K(K(A)) = K(A)$.

REMARK 12.10 *As Martin and Pollard, 1996 point out, the three conditions of Definition 12.9 can be replaced by the single condition:*

$$A \subseteq K(B) \Leftrightarrow K(A) \subseteq K(B).$$

The closure K *is called* algebraic *if in addition we have:*

$$K(A) = \bigcup \{K(B) \mid B \text{ is a finite subset of } A\}.$$

Closure systems are of course ubiquitous, and at this generality one would expect that they have little specifically geometrical content. Yet (abstract) *convexity structures* as studied particularly by van de Vel, 1993, which we shall consider in Sec. 7.1, are in effect just algebraic closure systems. As an abstraction of affine structure we have the *matroid*: a species of closure system which will be considered later in some detail.

A striking difference between closure systems and Čech closure spaces is that the former always have enough closed sets (sets A such that $A = K(A)$) to determine the space completely. Indeed, as is well-known, we can alternatively define a closure system to be a set X together with a collection $\mathcal{C}$ of ("closed") subsets satisfying:

- the intersection of any set of closed sets is closed.

A generalization of this situation will be discussed in a moment.

As we shall see in Sec. 7, abstract convexity structures are not entirely suited to our purpose: they are not well adapted to the discrete situation. Here one could think of incorporating graph structure into the convexity:

DEFINITION 12.11 *A* graph convexity *is an algebraic closure system defined over the vertices of a graph G, such that each closed set induces a connected subgraph.*

Numerous such "convexities" have been studied by graph theorists (Duchet and Meyniel, 1983; Duchet, 1988); for example, a set A of vertices of a connected graph G is *geodesic convex* if, for each pair u, v of vertices, every vertex lying on a shortest path between u, v is also in A. Again, in any Helly graph we have the "neighbourhood convexity", consisting of those sets of vertices which are intersections of balls (Tsaur and Smyth, 2004). On the whole, however, graph structure of itself, augmented with abstract convexity, does not sustain geometrically adequate notions of discrete convexity. More elaborate structure is needed: see Sec. 7, 9 below.

Returning to closures in general, we note that the collection $\mathcal{C}$ of closed subsets of a closure system (X, K) is a complete lattice, since it possesses arbitrary meets. We may say that it is a complete lattice of subsets of X. What this means is that the ordering is the subset order, and that $\mathcal{C}$ is a poset possessing (arbitrary) joins and meets. The lattice operations, however, are not required to agree with those of $\mathcal{P}(X)$ (although, in the case of a closure system, the meets of course agree).

Less familiar than this is the fact that we can relax the axioms of a closure system, and still obtain a complete lattice of "closed" sets.

DEFINITION 12.12 *A* weak closure system *is a set X together with an operator K which is monotonic and idempotent.*

PROPOSITION 12.13 *In any weak closure system (X, K), the closed sets (that is, fixed points of K) form a complete lattice under set inclusion.*

Proof Let $(C_i)_{i \in I}$ be a family of closed sets. Define

$$W = K(\bigcup_i C_i).$$

Then we have:

1 W is closed.

2 $C_i \subseteq W$ for each i, since $C_i = K(C_i) \subseteq W$.

3 $W \subseteq Z$ for any closed upper bound Z of the family (C_i); indeed $W \subseteq K(Z) = Z$. Thus W is the join of the family (C_i).

QED

We shall see a substantial example of this in a moment (regular sets).

4.2 Lattice of regular sets of a closure space.

For our main example of (weak) closure systems, we begin by recalling the notion of *regular* open (or closed) sets in topology. These sets have often been used for modeling spatial logics such as RCC. We observe that the notion can be generalized to (Čech) closure spaces:

DEFINITION 12.14 *Let* (X, Cl) *be a closure space, and* $A \subseteq X$ *an interior set (that is,* $A = \mathrm{Int}(B)$ *for some* $B \subseteq X$*). Then* A *is a* regular *interior set if* $A = \mathrm{Int}(\mathrm{Cl}(A))$*. Regular closure sets: defined similarly.*

It is well-known (see for example Birkhoff, 1948) that, in any topological space, the regular open sets (likewise, regular closed sets) form a complete Boolean algebra. This is by no means the case in arbitrary Čech closure spaces. In order to obtain a satisfactory algebra of regular sets, we shall restrict the closure spaces considered via:

DEFINITION 12.15 *A closure space* (X, Cl) *is* weakly regular *if, for any interior set* $A \subseteq X$*, the closure* $\mathrm{Cl}(A)$ *is a neighbourhood of* A*.*

Every topological space is trivially weakly regular. With regard to non-topological closure spaces, we have the following easy result:

PROPOSITION 12.16 *Every tolerance space* $(G, \sim)$ *is, as a closure space, weakly regular.*

The term "weakly regular" (we would have preferred "semi-regular", but that already has another sense: Engelking, 1989) has been chosen because, as we shall see in a moment, the closure spaces satisfying this condition have a rich structure of regular sets. It may be objected that the term is ill-chosen, as there is no special connection with regular topological spaces. Note, however, the following:

DEFINITION 12.17 (ČECH, 1966) *A closure space* (X, Cl) *is* regular *if, for every point* x *and neighbourhood* U*, there is a neighbourhood* V *of* x *such that* $\mathrm{Cl}(V) \subseteq U$*.*

PROPOSITION 12.18 *Every regular closure space is weakly regular.*

It is easy to see that weak regularity can be expressed in terms of the interior/closure operator, without explicit mention of points, by: $\mathrm{Int}A \subseteq \mathrm{Int} \circ \mathrm{Cl} \circ \mathrm{Int}A$.

For the definition of weak regularity to be meaningful, and for the following proposition to hold, we actually only need that Cl is a monotonic operator (with Int its dual). However for our intended application, namely Theorem 12.20 (see part 4 of the theorem), we require a Čech closure space, so we shall for simplicity work in that context. Abbreviate the operators $\mathrm{Int} \circ \mathrm{Cl}$, $\mathrm{Cl} \circ \mathrm{Int}$ by J, K respectively.

PROPOSITION 12.19 *Consider the following properties of a given closure space* (X, Cl)*:*

1. $\mathrm{Int}\ A \subseteq \mathrm{Int}\ \mathrm{Cl}\ \mathrm{Int}(A)$ *(i.e. X is weakly regular);*
2. $\mathrm{Cl}\ \mathrm{Int}\ \mathrm{Cl}(A) \subseteq \mathrm{Cl}\ A$*;*
3. *The operator K is idempotent;*
4. *The operator J is idempotent.*

Then: (1) $\Leftrightarrow$ (2), (3) $\Leftrightarrow$ (4), and (1) $\Rightarrow$ (3).

Proof Assume (1). By taking complements ($\neg$) we obtain:

$$\neg\mathrm{Int}\neg\neg\mathrm{Cl}\neg\neg\mathrm{Int}\neg(\neg A) \subseteq \neg\mathrm{Int}\neg(\neg A).$$

That is,

$$\mathrm{Cl}\ \mathrm{Int}\ \mathrm{Cl}(\neg A) \subseteq \mathrm{Cl}(\neg A).$$

Since A is arbitrary, we deduce (2). (2) $\Rightarrow$ (1) is similar. A similar manipulation with complements shows that (3) $\Leftrightarrow$ (4).

Now assume (1) (or, equivalently, (2)). We obtain $J \circ J \leq J$ by applying Int to each side of (2). On the other hand, we get $J \leq J \circ J$ by substituting $\mathrm{Cl}(B)$ for A in (1). QED

From Proposition 12.19 we see that, in any weakly regular space, the operators $J = \mathrm{Int} \circ \mathrm{Cl}$ and $K = \mathrm{Cl} \circ \mathrm{Int}$ are weak closure operators (that is, each is monotonic and idempotent, cf. Definition 12.12). In the following theorem we also make use of the operator $*$, defined by: $A^* = \mathrm{Int}(X - A)$. Trivially, we have in any space that $J(A) = A^{**}$; and if A is regular interior (that is, $A = A^{**}$), then A^* is also regular interior.

THEOREM 12.20 *Let $\mathcal{R}$ be the collection of regular interior sets of the weakly regular space* (X, Cl)*. We view $(\mathcal{R}, \subseteq)$ as a poset, and * (see the preceding remarks) as an operator $^* : \mathcal{R} \to \mathcal{R}$. Then we have:*

1 *$\mathcal{R}$ is a complete lattice;*

2 $A \leq B \Rightarrow B^* \leq A^*$*;*

3 $A^{**} = A$*;*

4 $A \vee A^* = X$ *and* $A \wedge A^* = \emptyset$*.*

Proof

1 By Proposition 12.13 applied to the weak closure system (X, J).

2 By monotonicity of Int.

3 A is a regular interior set.

4 $\mathrm{Cl}(A \cup A^*) = \mathrm{Cl}(A) \cup \mathrm{Cl}(A^*) = X$. Hence X is the only regular set containing both A and A^* as subsets.

QED

The clauses (2)–(4) of the preceding Theorem assert that * is an *orthocomplement* for the bounded poset $\mathcal{R}$. Together with (1), the statement is that $\mathcal{R}$ is a *complete ortholattice*.

EXAMPLE 12.21 *Let $(G, \sim)$ be a reflexive graph. As usual, we regard G as a closure space. It is not difficult to see that in this closure space, every closure set (that is, union of discs, or 1-neighbourhoods) is regular. Dually, every interior set is regular. Note also that, if $(G, \perp)$ is the complement (orthogonality) graph, we can write $A^\perp$ for $\mathrm{Int}(G - A)$, so that we have the expression $A^{\perp\perp}$ for $J(A)$ $(= \mathrm{Int} \circ \mathrm{Cl}(A))$. Also, $A \subseteq A^{\perp\perp}$; it is convenient to refer to $A^{\perp\perp}$ as the* orthoclosure *of A. The notation $O\mathcal{L}(G)$ will sometimes be used for the lattice of orthoclosed subsets of G.*

It may happen (as in Sec. 5.1) that an irreflexive graph $(G, \perp)$ is the primary object of study. In that case we still take the orthoclosed subsets to be those of the form $J(A) = A^{\perp\perp}$, while mention of Int, Cl, etc., will be taken to refer to the reflexive graph co-G.

These orthoclosed (alias regular interior) sets are extremely important for our subsequent development. As already mentioned in the Introduction, they provide our notion of *region*: see especially Sec. 7 for the details. In calculating the orthoclosure in a given $(G, \sim)$, a good way to think about this operation is as follows. To extend the set $A \subseteq G$ to $A^{\perp\perp}$, we add to A every vertex a such that each vertex adjacent to a is already adjacent to some vertex of A. In more geometrical language, the criterion is that $a \in A^{\perp\perp}$ if the 1-neighbourhood of a is contained in the 1-neighbourhood of A.

Most of the preceding material on regular interior sets applies, mutatis mutandis, to regular closure sets. In fact, since in any closure space X a set A is regular interior iff $X - A$ is regular closure, the inverse of the complete ortholattice $\mathcal{R}$ (Theorem 12.20) is isomorphic to the complete ortholattice of regular closure sets. We note that J. L. Bell (Bell, 1986), in his approach to quantum logic, has worked with the lattice of (regular) closure sets of a tolerance space (or *proximity* space in Bell's terminology). We shall take up the topic of quantum logic in the next section. We shall work there with the orthoclosure as defined above. This is more standard than the (Bell's) approach in terms of closure sets, and is also more convenient for our intended geometrical applications.

5. Extended examples

This section is mainly concerned with the modelling of various physical and spatial logics by lattices of subsets of a closure space (usually a graph). We pursue this theme via three extended examples.

The first of our three examples is concerned with developments of (so-called) quantum logic. The reader may wonder why so much attention is given to a subject so far removed from our main topic. In defence we would mention that the developments we are concerned with involve little that is specific to quantum theory. Rather, what is involved is a general logic of "tests", or "properties", the underlying structures of which (we shall argue) are of considerable significance for spatial reasoning.

5.1 Quantum structures

As is well known, Birkhoff and von Neumann, 1936 argued that the strangeness of quantum mechanics derived, in part, from the fact that its "logic" was non-classical: specifically, the distributive law failed. Less well known is the extensive development which these ideas have enjoyed from the 1950s onwards. We are concerned here with one of the main strands of this work, the "empirical logic" (logic of experiment) of D.J.Foulis and C.H. Randell. (Rather than cite a huge number of original papers, we mention here a couple of survey papers and a useful textbook: Cohen, 1989; Foulis, 1999; Coecke et al., 2000).

The departure point of the Foulis-Randell theory is an extremely simple definition:

DEFINITION 12.22 *A* test space *is a collection $\mathcal{A}$ of pairwise incomparable non-empty sets. Elements of $\mathcal{A}$ are called* tests. *Elements of tests are called* outcomes, *and the* outcome space *is $X = \bigcup \mathcal{A}$. The space is said to be* classical *if it has only one test. Finally, an* event *is a subset of a test.*

In the literature, variations on a thought experiment (due to Foulis) involving a "firefly box" are often used as illustrations of non-classical test spaces. The basic firefly "experiment" goes as follows:

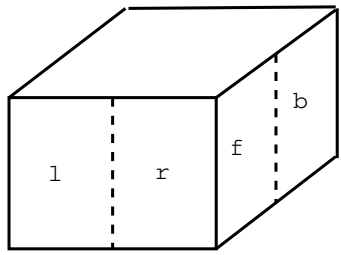


Figure 12.2. Firefly box.

EXAMPLE 12.23 *A rectangular box has two translucent windows, one on the front and one on a (specified) side; the other four sides of the box are opaque. At any given moment, the firefly's light is either on or off. This, we assume, can be detected by looking (directly) at either the front or the side. Moreover, by looking directly at the front window when the light is on, one can tell whether the firefly is in the left (l) or right (r) half of the box. Likewise, by looking directly at the side window when the light is on, one can tell whether the firefly is in the front (f) or back (b) half of the box. When the light is off, nothing (n) is seen, whether we look at the front or the side of the box. (It is assumed that the box has no partitions or baffles behind which the insect can hide, so that, if nothing is seen when looking at the front, then nothing would have been seen had one looked at the side, and vice versa.)*

With this set-up, then, we have two "experimental procedures" F and S. Procedure F is conducted by looking directly at the front of the box and recording l,r, *or* n *according to what, if anything, is seen. In procedure S one looks directly at the side of the box and records, as the case may be,* f, b *or* n.

Abstracting from this description, we have a test space consisting of the two tests $\{l, r, n\}$ *and* $\{f, b, n\}$. *The outcome space is* $\{l, r, n, f, b\}$.

As a somewhat less whimsical example we have:

EXAMPLE 12.24 *Quantum test space.* $\mathcal{A}$ *is the collection of all orthonormal bases of a Hilbert space* H. *Here, the outcome space* X *is the unit sphere of* H.

The third type of example we consider here is given as follows:

EXAMPLE 12.25 *Partition manual. A test space in which the tests form a partition (not just a cover) is called a* partition manual. *The* spin manual *is the special case of this in which each test has cardinality two. This name is, in*

part, a reference to the case of the "spin-1/2" particle, typically an electron, which, for each angle of measurement, has the two possible outcomes: spin up, or spin down.

Clearly, the notion of a test space is very general: just an irredundant cover of the set X. (Sometimes the definition is weakened further, to an arbitrary cover of X: Wilce, 2004.) We shall consider in a moment certain conditions that serve to make the notion more restrictive.

Any test space $(X, \mathcal{A})$ gives rise to an orthogonality graph $(X, \perp)$ via:

$$x \perp y \;\Leftrightarrow\; x, y \; are \; distinct \; elements \; of \; some \; test \; in \; \mathcal{A}.$$

This in turn gives a closure system (see Sec. 4), with closure operation given by:

$$A \mapsto J(A) = A^{\perp\perp}.$$

It is easy to check that in the Hilbert space example this simple definition does indeed give, as closed sets, the closed (linear) subspaces. In the spin manual, or any partition manual, closure is the identity on proper subsets of tests, while every other set has X as its closure.

A family $\{A_i\}$ of events of the test space $(X, \mathcal{A})$ is said to be *compatible* if its members are all contained in a common test. Disjoint compatible events are said to be orthogonal. Notation: $A \perp B$. (Notice that, so far, we are not entitled to identify this notion with orthogonality as applied to subsets of the graph $(X, \perp)$.) If $A \perp B$ and $A \cup B$ is actually a test, A and B are called *(orthogonal) complements*. Notation: $AocB$. Finally, the events A, B are said to be *perspective* (written $A \sim B$) if they possess a common complement.

DEFINITION 12.26 *The test space $(X, \mathcal{A})$ is said to be* algebraic *if, for any pair of perspective events A, B, any orthogonal complement of A is also an orthogonal complement of B.*

The idea of this definition is that perspective events can be considered as, in a sense, logically equivalent. For example, in the firefly experiment, we have that the events $\{l, r\}$ and $\{f, b\}$ are perspective. Each of these events occurs (under the appropriate test) if and only if the firefly's light is on, this being conveyed by the fact that their common complement is the event $\{n\}$. Thus, in a coherent experimental set-up, events which share (locally) an orthogonal complement should have the same complements everywhere: the test space should be algebraic.

Notice that any two tests of a test space are perspective (they have the empty set as complement). It is not difficult to show that, in the "logic of experimental propositions" of an algebraic test space, any test will appear as True (the top element of the lattice, assuming that we can construct an appropriate lattice).

Mathematically, a test space is nothing but a simplicial complex, where the events are the simplices and the tests are the maximal simplices. As such, a test space $(X, \mathcal{A})$ is graph-like (Sec. 2) if and only if every subset of X whose elements are pairwise orthogonal is an event. The assumption that this holds is (a strong form of) what is called *orthocoherence* in the quantum structures literature. The assumption was often made, more-or-less explicitly, in the older literature (say, from the 1950s to the 1970s), but is generally avoided today. However that may be, it is easy to see that the examples of test spaces given above (firefly, Hilbert space, partition manual) are all graph-like. Notice in particular that the spin manual having n tests is represented as the complete n-partite graph $K(2, 2, \ldots, 2)$—a very interesting graph to which we shall return more than once in the sequel. Moreover the example related to region geometry, with which we shall be concerned later, is graph-like. We shall adopt this assumption in the remainder of our discussion of test spaces and quantum logic. Thus, a test space is henceforth taken to be an orthogonality graph, in which the events are the cliques, and the tests are the maximal cliques.

In studying a test space $(G, \perp)$, we apply our previous work (see Theorem 12.20) on the ortholattice $O\mathcal{L}(G)$ of orthoclosed (that is, regular interior) sets of G. The closed sets we are concerned with now are those which are closures of cliques. These in general form a proper subset of $O\mathcal{L}(G)$. For example, in the graph co-C_5, any non-orthogonal pair of vertices is closed, but is not the closure of a clique. In fact, we shall need to consider *orthoposets* that are not necessarily lattices. An orthoposet is a poset P satisfying conditions (2)–(4) of Theorem 12.20; that is, P has the greatest element 1 and least element 0, and is equipped with an order-reversing, involutory unary operation * such that, for any element x, the join $x \vee x^*$ and the meet $x \wedge x^*$ exist and are equal to 1,0 respectively. A desirable property of our test space G is that the poset EC ("event-closures") of closures of cliques is an orthoposet, and indeed a suborthoposet of R. (Thus co-C_5 does not give a well-behaved test space; on the other hand, co-C_6 is satisfactory on this count.) Recall now the idea that perspective events may be considered logically equivalent. An obvious problem with this is that we have not shown that perspectivity is an equivalence relation. This property will need to hold if we are to have a satisfactory "logic" of events, and a good way to ensure it is to ask that two events be perspective if and only if they have the same closure. If this holds, we will be able to identify *propositions* with event-closures, rather than having to deal with equivalence-classes of events. One might guess (correctly, as we shall see) that this feature is closely related to whether the test space is algebraic.

We come now to an extremely important notion: orthomodularity (of a poset, particularly of a lattice). This can be formulated in many equivalent ways, and we shall use a formulation which is not usually taken as the definition, but is convenient for our purposes. Note first that we say that two elements x, y of

an orthoposet $(P, \leq, ^*)$ are *orthogonal* if $x \leq y^*$. (This relation is, of course, symmetric.)

DEFINITION 12.27 *Let $(P, \leq, 0, 1, ^*)$ be an orthoposet. We say that P is* orthomodular *provided that, whenever x, y are orthogonal, we have:*

$$y \vee x = 1 \Rightarrow y = x^*.$$

Given the element x of the orthoposet P we may refer, somewhat ambiguously, to an element y orthogonal to x whose join with x is (defined and) equal to 1 as "an orthocomplement" of x. Then orthomodularity is the condition that orthocomplements are unique.

Lastly (in preparation for our main result) we define a property of graphs which may appear somewhat technical, but is in fact of considerable geometric and logical significance:

DEFINITION 12.28 *An orthogonality graph $(G, \perp)$ is said to be* Dacey *if, for any maximal clique C and vertices x, y such that $C \subseteq x^\perp \cup y^\perp$, we have $x \perp y$).*

For a detailed study of Dacey graphs, see Sumner, 1974. The Dacey graphs encountered in applications, *other* than in quantum logic, are usually hereditarily Dacey (meaning that every induced subgraph is Dacey). For a discussion, with several characterizations, of hereditarily Dacey graphs, see Brandstädt et al., 1999.

We can now state:

THEOREM 12.29 *Let $(G, \perp)$ be an orthogonality graph, R its lattice of orthoclosed subsets, and $P(= EC)$ the poset of event-closures (i.e. closures of cliques) in G. Then the following are equivalent:*

1 *G is Dacey.*

2 *Let A, B be cliques such that $AocB$. If $K = J(A)$ (i.e. $A^{\perp\perp}$), then $K^\perp = J(B)$.*

3 *P is a sub-orthoposet of R, and each element K of P is the closure of every clique which is maximal in K.*

4 *P is an orthomodular sub-orthoposet of R.*

5 *Two cliques are perspective if and only if they have the same closure.*

6 *As a test space, G is algebraic.*

Proof (1) $\Rightarrow$ (2). Under the stated conditions on A, B, K, we evidently have $J(B) \subseteq K^\perp$. Suppose that $y \in K^\perp$ (so that $y \perp A$). Let x be an arbitrary

vertex orthogonal to B. By the Dacey condition for G, we have $x \perp y$. This shows that $y \in J(B)$.

(2) $\Rightarrow$ (3). Assume (2). Suppose that $K = J(A)$, A a clique. Extend A to a maximal clique $A \cup B$, so that $AocB$. Then by (2), $K^{\perp} = J(B)$. (Thus P is a sub-orthoposet.) Now let A' be any clique maximal in K. It is easily checked that $A'ocB$. Since $K^{\perp} = J(B)$, it follows by (2) that $K = K^{\perp\perp} = J(A')$.

(3) $\Rightarrow$ (4). Assume (3). We have to show that P is orthomodular. Take $K \in P$, where $K = J(A)$, A a clique. An element L of P orthogonal to K has the form $J(B)$, where $B \perp A$. Moreover, if $K \vee L = G$, we must have $BocA$ (since, for any x orthogonal to $A \vee B$, we would have $x \notin J(A \cup B)$). Thus if L is an orthocomplement of K, we indeed have $L = K^{\perp}$.

(4) $\Rightarrow$ (5). Notice that two cliques having the same closure are necessarily perspective (without the need to assume (4)). For if $J(A) = J(B)$, and $AocC$, then we know that $B \perp C$ and also that the clique $B \cup C$ is maximal (since a vertex x orthogonal to $B \cup C$ would be orthogonal to $J(B) \cup C$, and therefore to $A \cup C$).

Now assume (4). Suppose that we have (perspective) cliques A, B, with $AocC$ and $BocC$. Consider $J(C)$. Since P is a sub-orthoposet, the orthocomplement (in $O\mathcal{L}(G)$) K of $J(C)$ is an element of P. Now $J(A)$ is orthogonal to $J(C)$ and $J(A) \vee J(C) = G$; so by orthomodularity, $J(A) = K$. Likewise, $J(B) = K$.

(5) $\Rightarrow$ (6). Assume that, whenever two cliques are perspective, they have the same closure. Let A, B be cliques with $AocC$ and $BocC$. If also $AocD$, then $J(B)(= J(A)) \perp D$. As before, the clique $B \cup D$ is maximal; thus $BocD$.

(6) $\Rightarrow$ (1). Suppose that (as a test space) G is algebraic. Let A, B be cliques, with $A \cup B$ maximal, and x, y vertices such that $x \perp B$ and $y \perp A$. Choose a clique B' such that $Aoc\{y\} \cup B'$, and likewise a clique A' such that $A' \cup \{x\}ocB$. Thus we have $AocB$, and also $Aoc\{y\} \cup B'$. But $BocA' \cup \{x\}$, and so by algebraicity we get $A' \cup \{x\}oc\{y\} \cup B'$. In particular, $x \perp y$. QED

A "quantum logic" is sometimes defined to be an orthomodular poset (Kalmbach, 1983; Pták and Pulmannová, 1991). More often, perhaps, the structure is required to be an orthomodular lattice. The examples mentioned earlier in this subsection all yield orthomodular lattices rather than just posets. Several further refinements and generalizations of these structures are currently studied (Coecke et al., 2000; Dvurečenskij and Pulmannová, 2000). We do not here comment on these developments or their possible experimental justification. But we shall find that the orthomodular posets and lattices are of considerable significance for our geometric enquiry (Sec. 7).

We conclude this subsection with a characterization, due to Foulis and Randell, 1971, of "complete orthomodular spaces", that is, of graphs $(G, \perp)$ such that the lattice of orthoclosed subsets of G is orthomodular:

PROPOSITION 12.30 *G is a complete orthomodular space if and only if every closed set K of vertices is the closure of every maximal clique contained in K.*

Proof If the stated condition holds, then $R(= O\mathcal{L}(G))$ coincides with $P(= EC)$; hence by Theorem 12.29 R is orthomodular.

Suppose that R is orthomodular. Let $K \in R$, and let A be a maximal clique in K. Then $J(A) \perp K^{\perp}$, and moreover $J(A) \vee K^{\perp} = G$ (since clearly there cannot be a vertex orthogonal to both $K^{\perp}$ and $J(A)$). Hence $K = J(A)$. QED

5.2 Region Connection Calculus

Connection theory, also known as mereotopolgoy, is an approach to topology (and possibly geometry) in which regions rather than points are taken as the basic entities. Besides the set of regions, the main primitive is a two-place relation of "connection" between regions. A treatment of space(-time) on the basis of these primitives was proposed by Whitehead and others, on philosophical grounds, in the early XXth century.

Connection theory has enjoyed a revival in recent years, especially as it has become popular as a framework for spatial reasoning in computer science. It is, indeed, an active area of research, and quite a large literature has developed. Our justification for treating it as a mere "example" is that, in its main lines of development, connection theory implies that space is continuous (or infinitely divisible). Only a few attempts have been made to adapt the theory so that it can have non-trivial discrete models. Following a brief review of the general theory, we shall consider one such proposal, due to J. Stell.

Articles by B. L. Clarke (Clarke, 1981; Clarke, 1985) are credited with sparking the contemporary interest in connection theory. With R a non-empty set (of non-null "regions") and C a binary relation ("connection") we have, according to Clarke, the following axioms:

A1 xCx ;

A2 $xCy \Rightarrow yCx$;

A3 $C(x) = C(y) \Rightarrow x = y$.

Here, $C(x)$ means $\{y \in R|\, xCy\}$. Thus, in our terminology, **A3** amounts to T_0-separation. Given a non-empty subset X of R, a region x is said to be the *fusion* of X provided that, for every region y, yCx if and only if yCz for some $z \in X$; that is, provided that

$$C(x) = \bigcup\{C(z)|\, z \in X\}.$$

Then the next axiom is:

A4 Every non-empty set of regions possesses a fusion.

From **A3** we see that the fusion of the non-empty set X , say FX, is uniquely defined. Moreover (as in Sec. 3 above) the relation $\leq$ defined by:

$$x \leq y \quad \Leftrightarrow \quad C(x) \subseteq C(y)$$

is a partial order.

One aim of connection theory is to provide a definition of "point" as a suitable set of regions. Clarke's definition has it that a point is a pairwise connected set of regions, having some aspects of a prime filter. We shall not go into the details of this definition, but in terms of it Clarke states his fifth axiom:

A5 $xCy \quad \Rightarrow$ there exists a point P such that $\{x, y\} \subseteq P$.

Now Biacino and Gerla (Biacino and Gerla, 1991) have shown that a model of Clarke's axiom system **A1–A5** is, after adjunction of the null region 0, nothing but a complete Boolean algebra (under the ordering $\leq$), in which we have:

join is fusion;
complement is given by $\neg x = F\{z| \sim zCx\}$;
connection is overlap: xCy iff $\exists z.z \leq x \wedge z \leq y$ (written xOy) ;
points are prime filters.

Clearly, such a theory is unsuited to our needs. Simple graph models cannot be presented as Boolean algebras (without additional structure). By a "simple graph model" we mean a model in which the regions are subsets of some graph G, such that G (perhaps up to T_0-equivalence) can be recovered from the algebra. Even worse follows when Clarke's sixth axiom (i.e. Axiom A2.1$'$ of Clarke, 1985) is added: as Biacino and Gerla show, the addition of this axiom is equivalent to requiring the Boolean algebra to be atomless, thus ruling out any kind of discrete model.

Despite these unfortunate features, our review of Clarke's theory has not been a waste of time. As we shall see later, the restriction of this theory to the Axioms **A1–A4** can be very interesting for us.

RCC (Region Connection Calculus: Randell et al., 1992; Cohn et al., 1997) is inspired by Clarke's theory, but does away with the second order aspect (Axiom **A3**) and the reference to points (**A5**). As a purely first order theory of regions, it can claim to have advantages both methodological and practical (as a basis for automated reasoning).

We shall not here consider the original RCC axioms, but rather an essentially equivalent algebraic formulation due to Stell, 2000:

DEFINITION 12.31 *Let A be a Boolean algebra. Take R ("regions") to be the set of non-null elements of A; and take R_- (the proper regions) to be $R - \{\top\}$. We say that $(A;\ C)$, where C is a binary ("connection") relation on A, is a (Boolean)* connection algebra *provided:*

(B1) *C is symmetric, and its restriction to R is reflexive;*

(B2) *Every proper region is connected with its complement;*

(B3) *For non-null x, y, z : $C(x, y \vee z) \Leftrightarrow C(x, y)$ or $C(x, z)$;*

(B4) $\forall x \in R_- \ \exists y \in R. \neg\, xCy.$

The original RCC lacks the null element 0 (or $\perp$). This has had to be added to obtain a Boolean algebra, but the axioms **B1–B4** are silent as to the connection properties of $\perp$. Stell has pointed out (private communication) that this can be remedied by removing the restriction to non-null regions in Axiom **B3**. Assuming this slight modification, we get that $\perp$ is not connected with any region. In the following remarks we assume that all regions mentioned are non-null. An easy deduction from **B1** and **B3** is that

$$w\ O\ z \ \Rightarrow\ w\ C\ z.$$

Now if x, y are regions such that $\neg(y \leq x)$, then x' overlaps y in the region $y - x$, and so $x'\ C\ y$. By contraposition we get:

$$\neg\, yCx' \ \Rightarrow\ y \leq x. \tag{12.1}$$

Let x be a proper region. By **B2** we have xCx'. By **B4** there exists a region y not connected with x', and by (12.1) any such y must be strictly less than x (that is, $y \leq x$ and $y \neq x$). Repeating the step, we get an infinite decreasing sequence of regions starting from x.

Clearly, the fact that the lattice of regions is a Boolean algebra has been used heavily in this argument for the infinite divisibility of regions. Perhaps, then, a way to modify RCC so as to admit discrete models may be found by relaxing the assumption that the regions form a Boolean algebra.

This indeed is the approach of Roy and Stell, 2002. As a first step, observe that the RCC axioms not only preclude discrete models, but any models (even if infinite) in which there occur atomic regions. But the argument (using **B4**), that a region x cannot be atomic, would fail if it were possible for the complement of x to be $\top$. For this possibility to be achieved, we may think about replacing the Boolean complement with the dual pseudo-complement:

DEFINITION 12.32 *Let L be a lattice with top element $\top$. If $a \in L$, an element a' of L is the* dual pseudo-complement *of a provided that a' is the least element of L whose join with a is $\top$; that is*

$$\forall x \in L(a \vee x \ = \ \top \ \textit{iff}\ a' \leq x).$$

L will be called a dual p-algebra *if it is distributive, and the dual pseudo-complement a' exists for every $a \in L$.*

EXAMPLE 12.33 . *(A) The closed sets of any topological space X form a dual p-algebra. (The dual pseudo-complement of the closed set P is $Cl(X - P)$.)*

*(B) Let A be the dual p-algebra of closed sets of X, and define connection to mean overlap. Then, trivially, axioms **B1** and **B3** are satisfied. Moreover, **B2** holds if X is connected (in fact, **B2** is equivalent to the connectedness of X). Finally, **B4** is satisfied in certain cases, for example if X is $\mathbb{R}$.*

With an example like $\mathbb{R}$ we have a dual p-algebra which satisfies all the axioms **B1**–**B4**, and in which every region (closed set) is a join of atomic regions. If, with Roy and Stell, we want the model as a whole to be finite, **B4** is problematic. For this reason, Roy and Stell drop **B4**, arriving at the notion of a *connection algebra* as a dual p-algebra satisfying **B1**–**B3**. Their main example is that of a cellular complex, viewed as a poset with Alexandroff topology (that is, with the lower sets taken as the closed sets, and thus as the regions).

This approach is not taken very far by Roy and Stell, and it is not clear whether such a weak version of RCC is really viable. Moreover, the reduction of connection to overlap, as in these closed set models, seems to be contrary to the spirit of connection theory.

Yet the idea of relaxing the Boolean algebra requirement is surely worth pursuing. After all, in the (by a long way) most developed theory of "regions", namely frame/locale theory, we have a Heyting algebra rather than a Boolean algebra. If we give up the attempt to present our region theory as a modified RCC, it is possible to describe a rich algebra of regions, which is non-trivial even in the case of finite "spaces" (specifically, graphs). This was shown by Stell and Worboys, 1997, in work which we briefly discuss in our third (and last) main Example.

5.3 Co-Heyting algebras

Heyting algebras are very well known as structures with which to develop point-free ("region-based") topology. Also interesting for geometry, although much more rarely discussed, are the dual Heyting (or co-Heyting) algebras. Lawvere, 1991 argued for the geometric significance of co-Heyting algebras, and the argument has recently been taken up by a number of authors, notably Stell and Worboys, 1997 (in connection with region geometry), and Pagliani, 1998 (in the context of rough sets theory). We begin with the formal definition.

DEFINITION 12.34 *Let L be a bounded distributive lattice. If there is defined in L a binary operation ("implication") $\longrightarrow: L \times L \to L$ such that, for all $x, y, z \in L$,*

$$x \leq y \longrightarrow z \;\Leftrightarrow\; x \wedge y \leq z,$$

we say that $(L, \longrightarrow)$ *is a* Heyting algebra. *Dually, if in* L *we have a binary operation ("subtraction")* $\backslash$, *satisfying:*

$$y \backslash z \;\leq\; x \;\;\Leftrightarrow\;\; y \leq x \vee z,$$

then $(L, \backslash)$ *is a* co-Heyting *algebra. If* L *has both an implication* $\longrightarrow$ *and a subtraction* $\backslash$, $(L, \longrightarrow, \backslash)$ *is a* bi-Heyting *algebra.*

REMARK 12.35 *Evidently,* L *can have at most one implication (resp. subtraction) defined on it. Every Boolean algebra is bi-Heyting, since Boolean implication (* $\neg x \vee y$*) is both an implication and a subtraction in the sense of Definition 12.34. Indeed, a bi-Heyting algebra* $(L, \longrightarrow, \backslash)$ *is a Boolean algebra iff* $\longrightarrow$ *coincides with* $\backslash$.

If we think of the elements of L as regions then, informally, $y \longrightarrow z$ is the largest region (if it exists) whose intersection with y is contained in z. The standard example is that in which L is taken to be the collection of open sets of a topological space. Dually, $y \backslash z$ is the smallest region whose union with z contains y. One readily guesses (and verifies) that the closed sets of a topological space constitute a co-Heyting algebra.

We shall here be concerned with certain discrete spatial models, rather than general topological spaces. Moreover, for the existing applications we do not actually need the full binary operations, but only the corresponding (unary) complements. That is, we make use just of the *negation* $\neg : L \to L$, defined by:

$$\neg x \;\;=\;\; x \longrightarrow 0,$$

or (in the co-Heyting case) of the *supplement* $\sim : L \to L$:

$$\sim x \;\;=\;\; 1 \backslash x.$$

An advantage claimed for the co-Heyting (or bi-Heyting) approach is the simple algebraic treatment of *boundaries* which it permits. Let us see how this works in the important *graph* example (Lawvere, 1991; Reyes and Zolfaghari, 1996; Stell and Worboys, 1997). In this context, it is usual to consider a graph G to be a pair (V, E), where V, E are (disjoint) sets of vertices and edges, respectively, and each edge has either one or two vertices (thus multi-edges, as well as loops, are allowed). A graph $G' = (V', E')$ is a *subgraph* of G provided that $E' \subseteq E, V' \subseteq V$, and each edge $e \in E'$ has the same vertices in G' as it has in G. Let, now, L be the collection of subgraphs of a graph G. Clearly, L is a bounded distributive lattice. But complements are not so straightforward. If (V', E') is a subgraph of G, the "pairwise" complement $(V - V', E - E')$ is not necessarily a subgraph: an edge e belonging to $E' - E$ may have a vertex lying in V'. This is a symptom of the fact that L is not in general a Boolean algebra.

It is, however, a bi-Heyting algebra. (This may be verified directly; it is also an immediate consequence of Remark 3 to follow.) The subgraph (V', E') thus has a supplement. This is the smallest subgraph whose union with (V', E') is G. That is, $\sim (V', E')$ is the set $E - E'$ of edges, together with their end-points.

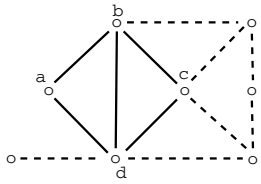


Figure 12.3. Region R.

Figure 12.4. Supplement, $\sim$ R.

Figure 12.5. Boundary.

This leads to the suggestion that the boundary of a "region" (subgraph) of a graph may be defined in the same way as is done in topology. The boundary of a closed subset C of a space X is by definition the intersection of C with the closure of $X - C$. In terms of co-Heyting algebra, this is $C \wedge \sim C$. The suggested definition is thus: the *boundary* of a graph region R is $R \wedge \sim R$. Thus, in Fig. 12.5, we get the vertices b, c, d (but no edges between distinct vertices).

Remark 1 By this definition, the boundary of a graph region is always "totally disconnected" (it contains no edges).

Remark 2 Reyes and Zolfaghari, 1996, work with digraphs. However it is easy to see that the orientation plays no part in the treatment of boundaries (or in the bi-Heyting structure generally). Nothing is lost by confining attention (as in Stell and Worboys, 1997) to undirected graphs.

Remark 3 It is easy to see that the co-Heyting algebra of subgraphs is a special case of the closed sets algebra. Given a graph $G = (V, E)$, let $X = V \cup E$. Then the subgraphs of G (more precisely, the subsets $V' \cup E'$ corresponding to subgraphs $(V' \cup E')$) are the closed sets of a topology on X. Spelling this out a little: order the set X by taking $x \leq y$ if $x = y$ or x is a vertex (end-point) of the edge y. Then the closed sets (in effect, the subgraphs) are simply the downward-closed subsets of X. Of course, we could equally well take these subsets as the open sets of a topology on X, which "explains" why the subgraphs of G give us a bi-Heyting algebra, not just a co-Heyting algebra.

The notions *boundary* and *dimension* are closely related. We expect that, in general, a k-dimensional region will have a $k-1$-dimensional boundary. From this point of view the preceding treatment of graph boundaries is somewhat disappointing: boundaries are always (on any reasonable definition of dimension) 0-dimensional. In topology, graphs are usually treated as one-dimensional "complexes", so it could be said that this result is only to be expected. But if we are hoping that graphs can provide us with reasonably general "discrete spatial models", a different treatment will be needed. This we attempt to provide in the next section.

Stell and Worboys, 1997, provide a second type of example of bi-Heyting algebras, involving pairs of sets. Specifically, let L be the collection of pairs (S, T) of subsets of a ground set A, such that $S \subseteq T$. L is made into a distributive lattice by defining the lattice operations componentwise from the corresponding operations over the subsets of A. For operations that are not monotonic in every argument more care is needed, due to the constraint which has to be respected. But it can be checked that the following formula defines subtraction in L:

$$(S_1, T_1) \setminus (S_2, T_2) \;=\; (S_1 \cap A \setminus T_1, (S_1 \cap A \setminus T_1) \;\cup\; (S_2 \cap A \setminus T_2)).$$

The dual formula gives implication.

One possible interpretation of this scheme has to do with vagueness: A pair (S, T) is thought of as a description of a region with indeterminate boundaries, where S consists of elements which are certainly in the region, while T has elements which may be in the region. This example appears to be of quite a different character than the previous one involving graphs. Yet it also is reducible to closed sets algebra. To see this (with some details left to the reader), observe that pairs of subsets of A may be identified with subsets of $A \times 2$ $(= \{(a, i) |\ a \in A, i \in \{0, 1\}\})$. Next: determine a partial order $\leq$ on $A \times 2$ such that the pairs $(S, T) \in L$ are thereby identified with the lower sets of $(A \times 2, \leq)$. In this way we get a "closed sets" algebra of the pairs.

The three extended examples in this section are intended to set the scene for our own positive proposals in the remaining sections. Note in particular the aspect of distributivity of the lattices involved. Quantum logic generally has non-distributive ortholattices. Studies in spatial reasoning on the other hand almost invariably assume distributivity, and often assume that a Boolean algebra is given. These distributivity assumptions are, we believe, at the source of difficulties which researchers in spatial reasoning have experienced in handling such notions as boundary and complement (especially when "discrete" models are intended). It is in this context that it can be beneficial to look at the work that has been done in quantum structures.

6. (Boundary and) dimension

As our next major topic, we consider dimension in discrete spaces. Dimension and boundary are closely related ideas. One of the main intuitions about dimension is that the dimension of a space should be one greater than that of the boundaries of portions of the space. The definition of "boundary" which we have seen in the previous section gives the result that all boundaries are totally disconnected. On this approach, every space will be assigned a dimension ≤ 1.

On the other hand, a too-literal adaptation of the topological boundary concept will (as we shall see) result in an implausibly high dimension for simple discrete spaces. Even the careful approach of Evako, 1994, Evako et al., 1996, which is the closest to that which we shall adopt here, results in uncomfortably "high" dimension for our basic spatial models.

We begin by reviewing some of the main desiderata for a theory of dimension. Just as in topology (Hurewicz and Wallman, 1948), we shall expect:

1 The dimension function ($\dim$) is a "topological" invariant.

2 The dimension of I_m^n (our discrete version of the n-cube) should be n.

3 Monotonicity. If X is a subspace (induced subgraph) of Y, $\dim(X) \leq \dim(Y)$.

Moreover we continue to adhere, as far as possible, to our Correspondence Principle. Specifically, we aim to provide a formulation that works well at least for tolerance spaces, and possibly for larger classes of graphs, and is "close enough" to standard topology. Still more convincing, no doubt, would be to provide an account at the level of closure spaces: a common generalizaton of topological and tolerance space definitions. As to whether this is possible, however, we shall have to leave for later research to determine.

Recall the definition of "small inductive dimension", $\dim$, in topology:

DEFINITION 12.36 *If X is the empty space,* $\dim(X) = -1$. *A non-empty space X has dimension $\leq n$ provided that* $\dim(x) \leq n$ *for every $x \in X$, where* $\dim(x) \leq n$ *means that x has arbitrarily small neighbourhoods with boundary of dimension $\leq n-1$.*

This is easy to calculate in spaces having Alexandroff topology, as we only have to consider the (unique) smallest neighbourhood $N(x)$ of x when computing $\dim(x)$.

EXAMPLE 12.37 *Let X be a finite poset with the Alexandroff (upper sets) topology. Then* $\dim(X) = h$, *where h is the length of (the longest chain in) X. Indeed, suppose that x is a maximal point of X. Then the least neighbourhood of x is $\{x\}$, and its boundary* $\mathrm{Cl}(\{x\}) - \mathrm{Int}(\{x\})$ *is the set of points strictly*

below x, which has length $h-1$. Moreover, every set of the form $\uparrow y$, $y \in X$ clearly has a boundary with length $\leq h-1$. An inductive argument yields the result, $\dim(X) = h$.

Suppose that we try to apply this formulation, unmodified, to a tolerance space G. Of course, we view G as a closure space, and $\mathrm{Cl}, \mathrm{Int}$ are interpreted accordingly. First, if G is a clique, then it is assigned the dimension 0 (the boundary of every neighbourhood is empty). Next, if G consists of some cliques joined in a row, as for example G' in Fig. 12.6, then we get dimension 1 (the boundary of each $N(x)$ is either a clique or empty). If the cliques are arranged in a "corner", however, as in G'' (Fig. 12.7) we get dimension 2, since some of the vertices of G'' have neighbourhoods with boundary G'. The purpose of this series of little examples is to lead up to the case $G = I_m^2$ $(m \geq 2)$. In fact, we evidently get $\dim(G) = 3$ in this case (instead of 2 as we intend).

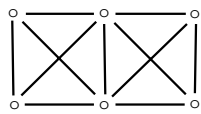


Figure 12.6. G': $dim = 1$.

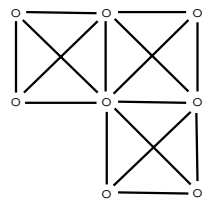


Figure 12.7. G'': $dim = 2$.

It is (it seems) easy to see why we are getting too high a dimension: as against the "boundaries" we encountered in the previous section (which were too thin), we now have boundaries which are too thick. A possible solution lies to hand: the approach taken in Evako et al., 1996. According to this, we ignore the topological definition of "boundary", and simply decide that the only boundary set that we need to consider, for a vertex $v \in G$, is $N(v) - \{v\}$. The digital n-sphere (Evako, 1994) provides a pleasing example. As presented by Evako, this is the complete n-partite graph $K(2, 2, \ldots, 2)$. Dually, it may be described as the result of deleting the edges between opposite pairs of vertices of the cube I_1^n. (These figures are combinatorially dual in the sense in which the regular octohedron and cube are combinatorially dual: vertices and faces are interchanged.)

This approach also, however, is not without problems. The dimension of the clique K_n now evaluates to $n-1$. As a result, we still get that $\dim(I_m^2) = 3$. Higher powers of I_m give an even worse result: since I_m^n contains 2^n-cliques, it has dimension at least 2^{n-1}.

We shall find that a slight modification of the definition of Evako et al., 1996, will give us all that we require, while leaving undisturbed their main

examples (the digital n-spheres, and surfaces generally). A first difficulty that we should address concerns the formula $N(v) - \{v\}$. This seems to be so far removed from the topological boundary formula that it must be forbidden by the Correspondence Principle. However, the discrepancy is not necessarily as great as it seems. It must be remembered that the closure space definition (for the required boundary) need not be syntactically the same as the standard topological definition, provided it reduces to the latter when the closure is idempotent. The topic is too speculative for us to go into details here, but there is evidence that at the closure space level the definition takes (in part) the form $\mathrm{Cl}(\mathrm{Int}(N)) - \mathrm{Int}(N)$, where N is the chosen neighbourhood of v. This is in fact the topological boundary, for open N, whilst it also corresponds closely to the formula $N(v) - \{v\}$ in the case of a tolerance space.

Although we shall retain the formula $N(v) - \{v\}$, we shall modify the condition which relates the dimension at a point to the dimension of the corresponding boundary set. The principle here is that the mere duplication of a point should not change the dimension of a space. More precisely, if $x, x' \in X$, with $x \equiv x'$ in the specialization order of X, then we should have $\dim(X) = \dim(X - \{x'\})$. For example, with X as the graph (tolerance space) G of Fig. 12.8,

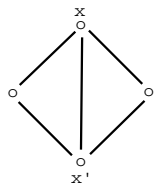


Figure 12.8. G.

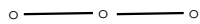


Figure 12.9. $G' = G \backslash x'$.

we get:

$$\dim(G) = \dim(G') = 1.$$

We thus arrive at the following definition. The notation here is that $\dim(G)$ is the dimension of G, while $\dim_G(x)$ is the dimension at the vertex x of G. Also we denote the punctured neighbourhood $N(x) - \{x\}$ by $N_0(x)$.

DEFINITION 12.38 *If G is empty,* $\dim(G) = -1$. *If G is non-empty,* $\dim(G) = \sup\{\dim_G(x) | x \in V(G)\}$. *For $x \in V(G)$, we put:*

$$\dim_G(x) = \begin{cases} \dim(N_0(x)), & \textit{if } \exists y \in N_0(x), N(x) \subseteq N(y) \\ \dim(N_0(x)) + 1, & \textit{otherwise.} \end{cases}$$

REMARK 12.39 *The first alternative in the expression for* $\dim_G(x)$ *in the preceding definition is usually expressed by saying that x is a* dominated *vertex.*

It means that, in the subspace $N(x)$, there is a vertex ($\in N_0(x)$) y such that $x \equiv_{N(x)} y$.

REMARK 12.40 *We shall henceforth assume that our tolerance spaces are finite dimensional (although this restriction can be avoided fairly easily if desired).*

In particular, if G is a clique, then all vertices of G are equivalent, and it follows that $\dim(G) = 0$. The dimension of cliques is particularly significant. It is easy to see, indeed, that under the definition of Evako et al., 1996, the dimension of G is the dimension of its largest clique, that is, $c(G) - 1$. Working instead with Definition 12.38, do we get any simple relation with the size of cliques? Let us consider those non-empty cliques of G which are intersections of (one or more) maximal cliques. As we shall see in a moment, there are grounds for considering these intersection cliques as the *cells* (or *faces*) of G. We shall denote the poset of these cells, ordered by inclusion, as $\mathrm{cell}(G)$. Recall that the *level* of an element x of a poset P (of bounded chain-length) is the length of a longest chain below x, and that the *length* of P is the level of an element of greatest level, that is, the length of a longest chain in P. We take the length of the empty poset to be -1.

THEOREM 12.41 *For any non-empty tolerance space G,*

$$\dim(G) = \mathrm{length}(\mathrm{cell}(G)).$$

Proof We prove by induction on $|V(G)|$ that, at each point x of G, $\dim_G(x)$ is the length of the poset $\mathrm{cell}_x(G)$ of those cells which contain x. This is obviously true if $|V(G)| = 1$. Suppose that $v \in V(G)$, with $N_0(v)$ non-empty. There are two cases. In the first case, v is dominated by $v' \in N_0(v)$. If C is any clique contained in $N_0(v)$, then $C \cup \{v'\}$ is also a clique contained in $N_0(v)$. Thus, given any chain K of cells in $N_0(v)$, we have a chain of cells containing v', of the same length as K, in $N_0(v)$. Since $v \equiv v'$, we have $\mathrm{cell}_v(N(v)) \cong \mathrm{cell}_{v'}(N_0(v)))$. By the induction hypothesis we can conclude that

$$\begin{aligned}\dim_G(v) &= \dim(N_0(v)) = \mathrm{length}(\mathrm{cell}_{v'}(N_0(v))) \\ &= \mathrm{length}(\mathrm{cell}_v(N(v))) = \mathrm{length}(\mathrm{cell}_v(G))\end{aligned}$$

(since a clique containing v is necessarily a subset of $N(v)$).

In the second case we have, for each $v' \in N_0(v)$, a $v'' \in N_0(v)$ such that $\neg(v'' \sim v')$. Any maximal clique containing v, v'' thus excludes v'. It follows that $\{v\}$ can be expressed as an intersection of maximal cliques; that is, $\{v\}$ is a cell. Let (invoking the induction hypothesis) u be a vertex in $N_0(v)$ such that $\dim(N_0(v)) = \mathrm{length}(\mathrm{cell}_u(N_0(v)))$. Further, let K be a maximal chain in

$\mathrm{cell}_u(N_0(v))$. By extending each cell $C \in K$ to $C \cup \{v\}$, and taking also the cell $\{v\}$, we obtain a (maximal) chain K' in $\mathrm{cell}_v(N(v))$, of length $\mathrm{length}(K)+1$. We conclude that, in this case,

$$\dim_G(v) = \dim(N_0(v)) + 1 = \mathrm{length}(\mathrm{cell}_v(N(v))) = \mathrm{length}(\mathrm{cell}_v(G)).$$

QED

We have already remarked that in many important cases our definition gives the same result as that of Evako et al., 1996. In particular this is the case for the closed surfaces as defined in that work:

DEFINITION 12.42 *A* 0-surface *is a disconnected graph (tolerance space) with exactly two vertices. For $n > 0$, an n-*surface *is a non-empty connected graph G such that, for each vertex v of G, $N_0(v)$ is an $(n-1)$-surface.*

In terms of topology, the surfaces defined here are closed surfaces, as opposed to surfaces with boundary. It is an easy exercise to show that every n-surface has dimension n at each of its points (vertices).

The "simplest" n-dimensional surface is the digital n-sphere (= n-crosspolytope as in Sec. 9 below; cf. also the "spin manual", Sec. 5.1): the complete $n+1$-partite graph $K(2, 2, \ldots, 2)$. We denote the digital n-sphere by DS^n. One way to make precise the claim about the simplicity of the sphere is given by the following;

PROPOSITION 12.43 *Let S be any n-surface. Then there exists a surjective morphism from S onto DS^n.*

Proof By induction on n. If $n = 0$, we have $S = DS^n$. Suppose then that $n = k > 0$, and assume for the induction that every $(k-1)$-surface can be mapped onto DS^{k-1}. Given the k-surface S, let p, q be arbitrarily chosen points of S, DS^k, respectively. The (punctured) neighbourhood $N_0(p)$ of p in S is a $(k-1)$-surface , and thus we have a surjection $e_{k-1} : N_0(p) \to DS^{k-1}$. we identify DS^{k-1} with the neighbourhood of q in DS^k. Denote by $-q$ the remaining point of DS^k (antipodal to q). Then the required surjection $e_k : S \to DS^k$ is defined by:

$$e_k(x) = \begin{cases} q, & \text{if } x = p \\ e_{k-1}(x) & \text{if } x \in N_0(p) \\ -q & \text{otherwise.} \end{cases}$$

This is a morphism, as p is not adjacent to any element of $S \setminus N(p)$. QED

In the next few Propositions we indicate some initial steps in the discrete theory of surfaces and dimension, focussing on decomposition results.

PROPOSITION 12.44 *Suppose that* $G = A * B$ *is a surface. Then each summand* A, B *is either empty or a surface.*

Proof By induction on $|G|$. The assertion is immediate for $|G| \leq 2$. Suppose then that $G' = A * B$ is a surface, with $|G'| = n > 2$, and assume (for the induction) that the assertion holds for all G with $|G| < n$. If one of the summands, say A, is empty, then there is nothing to prove, since in that case B is isomorphic with G', and is thus a surface. So assume that both summands are non-empty, and choose a vertex from either one of them, say $a \in A$. Then the (punctured) neighbourhood of a in G' is $N_0(a) * B$, where $N_0(a)$ is the neighbourhood of a in A. Since G' is a surface and $|G| > 2$, $N_0(a) * B$ is a surface. Hence (by the induction hypothesis) B is a surface. QED

COROLLARY 12.45 . *No cone (i.e. graph of the form* $1 * G$*) is a surface.*

PROPOSITION 12.46 *No surface is dismantlable.*

Proof In fact no tolerance space with a dominated vertex can be a surface. For suppose that v is a dominated vertex of the space G; say v is dominated by $u \in N_0(v)$. Then u is connected to every vertex of $N_0(v)$, so that $N_0(v)$ can be written as $u * (N_0(v) \setminus u)$. By the preceding Corollary, $N_0(v)$, and hence G, is not a surface. QED

The converse of Proposition 12.44 also holds. In fact we have the following (= Theorem 5 of Evako et al., 1996):

PROPOSITION 12.47 *If* G *is an* n*-surface and* H *is an* m*-surface, then* $G * H$ *is an* $(n + m + 1)$*-surface.*

As already mentioned, an important case in which we differ from Evako et al., 1996 is the "digital n-cube" I^n_m. It follows from the Product Theorem, to be given in a moment, that this receives the "correct" dimension. The same case serves to explain our terminology of "cells": by inspection we see that, at least in the interior of the digital cube, the cells are indeed the small cubes and their faces. Polytopes provide a further source of illuminating examples. In fact, given a polytope Π, we derive a graph $G(\Pi)$ by making each facet of Π into a maximal clique (that is, for each facet F, an edge is created for each pair of vertices of F). The graph $G(\Pi)$ and its cells thus constitute a combinatorial version of Π.

THEOREM 12.48 (PRODUCT THEOREM) *For any non-empty tolerance spaces G, H,*

$$\dim(G \times H) = \dim(G) + \dim(H).$$

Proof We note first that a subspace $S \subseteq G \times H$ is a maximal clique iff S is of the form $A \times B$, where A, B are maximal cliques in G, H respectively. Then, F is a face (of $G \times H$) contained in such S iff F is $A' \times B'$, where A', B' are faces of G, H contained in A, B respectively. Finally, note that if K is a maximal chain of cells in $G \times H$, with greatest element S, each link of K (connecting a $k+1$-cell in K to a k-cell contained in it) corresponds to a link in either A or B. In particular this applies if K is a chain of maximal length (below S); thus

$$\text{level}(S) = \text{level}(A) + \text{level}(B).$$

Applying this to an S of greatest level, we have the result. QED

Note. In taking account of Evako's approach, we have worked mainly with the systematic presentation in Evako et al., 1996. There are however indications in Evako, 1994, especially as seen in the Examples there, that, in the case of tolerance spaces, something closer to our Definition 12.38 was intended by Evako.

In the definition of closed surfaces (Definition 1.47), comparison with classical manifolds would suggest that, for each vertex v, $N_0(v)$ should be required to be an $(n-1)$-*sphere*, rather than just an $(n-1)$-surface. Indeed, something of this sort is generally done, both by Evako and by other authors. An accurate treatment of this theme would, however, require that we carefully consider what is to count as "a" digital n-sphere, rather than "the" (minimal) digital n-sphere discussed above. This in turn involves the question of the (homotopy-)equivalence of digital spaces: an extremely important topic, but one that is too complex to enter into here.

7. Discrete region geometry

7.1 Abstract convexity

To help motivate the approach which we take here to region geometry (and to convexity in particular), we begin this section with a brief discussion of the abstract convexity spaces which were mentioned in Sec. 4. In this context, an algebraic closure system will be renamed a *convexity structure*, the closed sets now being called *convex*. It is often convenient to impose the normalizing conditions that the empty set and all singletons are convex; let us assume that at least the first of these conditions is imposed. As regards notation, we shall for the moment use $Conv$ for the closure (i.e. convex hull) operation. (Later, a square bracket notation will be used.)

On this very slender basis one can define and study a range of notions important in convexity and polytope theory. Take an *open interval* to be any set of the form $Conv\{a, b\} \setminus \{a, b\}$. Notation: (a, b).

DEFINITION 12.49 *Let C be a convex set of the structure $(X, Conv)$. Then an* interior *point of C is any point $p \in C$ such that*

$$\forall x \in C \setminus p \; \exists y \in C.\, p \; \in \; (x, y).$$

Further, a face *of C is any convex subset F of C such that;*

$$\forall x, y \in C.\, (x, y) \cap F \neq \emptyset \; \Rightarrow \; x, y \in F.$$

Recall that in real analysis, the *relative interior* of a convex set $C \subseteq \mathbb{R}^n$, $relint(C)$, is defined as follows. First construct the affine hull $< C >$ of C in $\mathbb{R}^n$. Endow $< C >$ with the subspace topology derived from $\mathbb{R}^n$. Then take $relint(C)$ to be the topological interior of C in $< C >$. It happens that the abstract convexity approach enables one to define the relative interior, and other useful notions, with far less conceptual apparatus than is needed classically. In particular we have the following proposition, whose proof is left to the reader as an exercise:

PROPOSITION 12.50 *Let $C \subseteq \mathbb{R}^n$ be convex. Then $p \in relint(C)$ if and only if p is an interior point of C in the sense of Definition 12.49. Also, if P is a polytope in $\mathbb{R}^n$, Definition 12.49 identifies the faces of P in the usual sense.*

It is immediate from the definition that any intersection of faces of a convex set (in an arbitrary convexity structure) is again a face of C; thus we have a complete lattice of faces. A special case of this is the familiar face lattice of a polytope (in $\mathbb{R}^n$).

Any face of C other than C itself is called a *proper* face. Concerning proper faces and the set $\mathcal{I}(C)$ of interior points of C, we have the following very easy result:

PROPOSITION 12.51 *Every proper face of the convex set C is disjoint from $\mathcal{I}(C)$.*

Proof Let p be any element of the proper face F. Choose a point $x \in C \setminus F$. Then there does not exist any point $y \in C$ with $p \in (x, y)$ (for if such y exists, then by definition of a face, $x, y \in F$). Thus $p \notin \mathcal{I}(C)$. QED

One might hope for a stronger result here, namely that every point of C belongs either to $\mathcal{I}(C)$ or to a proper face (but not both). This, however, does not hold in an arbitrary convexity structure. To ensure that results such as this hold, one would need to impose additional axioms on the convexity structure involved;

see in particular (Coppel, 1998; van de Vel, 1993). We do not follow this path here, as the additional axioms include density properties which conflict with our desire for discrete models.

Given a convexity structure $(X, Conv)$ and a subset $Y \subseteq X$, we have the evident substructure convexity on Y, where the convex sets in Y are the sets $C \cap Y$, where C is convex in X. This gives us a plentiful supply of discrete convex structures.

EXAMPLE 12.52 *Let a, b, c be the vertices of a triangle in $\mathbb{R}^2$. Let p, q, r be the midpoints of the sides $bc, \ldots$, and and o the centroid, that is, the intersection of ap, bq, cr. Take the set T as $\{a, b, c, p, q, r, o\}$, and let $(T, Conv)$ be the convexity inherited from $\mathbb{R}^2$. We readily check that the sole interior point of T is o, and that the faces of T are as expected. However, if the point p is deleted, giving $(T \setminus p, Conv)$, the point o is no longer an interior point (nor does it belong to any proper face).*

For a more extreme illustration of the difficulties arising from not having "enough" points in a discrete model, consider the following:

EXAMPLE 12.53 *Let P be a polytope in some Euclidean space $\mathbb{R}^n$, and let T be the set of vertices of P. Consider the substructure $(T, Conv)$, with the inherited convexity. According to Definition 12.49, we have to regard every subset of T as a face of the polytope T.*

A face of a convex set is by definition a convex subset F such that no interval can "cross" F. Taking the polytope P of Example 12.53 to be a rectangle $abcd$, ac fails to be a face as it is crossed by the other diagonal bd. The question arises as to whether, in going to the discrete substructure in which we retain only the four vertices, there is some sense in which we can assert that the diagonals cross, even though they no longer have a point in common. (Intuitively, it appears to be so.) Our solution will in effect be to treat the notion, that two figures (especially polytopes) meet in their interiors, as a primitive—indeed, the main primitive - of our geometric theory. In terms of our work in previous sections, this is a connection relation which gives rise to various structures involving closure spaces and systems, as discussed above.

7.2 Cell Geometry Axioms

For our version of region geometry, then, we suppose that regions are built from "basic regions" thought of as cells (= polytopes). Moreover a connection relation, denoted $\approx$, is assumed to be defined over the cells, and thence over the regions. Formally, we take as a starting point the following definition:

DEFINITION 12.54 *A* cell geometry *is a pair* $(\mathcal{P}^+_{fin}(X), \approx)$*, where* $\approx$ *is a reflexive, symmetric relation defined on the non-empty finite subsets of a ground-set* X*, satisfying:*

(Add) $A \approx B,\ A' \approx B' \ \Rightarrow\ A \cup A' \approx B \cup B'$.

Instances of the relation $\approx$ *will sometimes be called* quasi-equations.

Remark. We require only additivity, rather than monotonicity ($A \approx B \subseteq B' \ \Rightarrow\ A \approx B'$), as we want to allow for the interpretation that $A \approx B$ means that the cells A, B meet in their relative interiors. For example, if B is (interpreted as) a closed interval $[xy]$ in $\mathbb{R}$, with $x < y$, then we have $\{x\} \approx \{x\}$ while $\{x\} \not\approx [xy]$, so that monotonicity fails.

Given a cell geometry $(\Sigma, \approx)$ we define a *region* to be an (ortho)closed set of cells. The cell geometry is here being viewed as a tolerance space. A simple example may illustrate the idea. Let X be some subset of intersection points (vertices) of a rectangular grid in $\mathbb{R}^2$, such as the set of nine vertices named in Fig. 12.10. In this case we have $2^9 - 1$ "cells", according to Definition 12.54. We adopt our standard interpretation of $\approx$, namely that the connection of two cells means the overlapping of their interiors. (N.B. In this context, "interior" is always taken in the sense of Definition 12.49, or equivalently the relative interior of conventional convex analysis.) In denoting cells, we shall often omit set brackets and commas. Thus we have, for example, $ag \approx bfh$, and $c \approx bdh$.

Let S be the set of "cells"

$$ab,\ abcd,\ b,\ bc,\ efh,\ e,\ bi;$$

see Fig. 12.10.

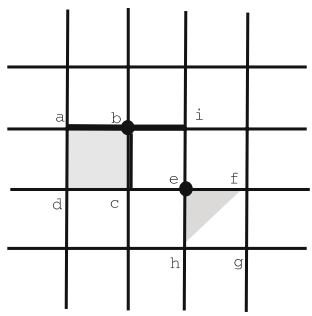


Figure 12.10. S : ab; $abcd$; b; bc; efh; e; bi.

In terms of the Euclidean model, S determines the region $\bigcup S$, and any cell C can be considered as a part of this region if it is covered (with respect to interiors) by $\bigcup S$. In terms of our formal cell geometry, however, we have to calculate

the orthoclosure $J(S)$ (in the notation of Sec. 4). Recall that $J(S) = S^{\perp\perp}$, and that we also have

$$A \in J(S) \Leftrightarrow N(A) \subseteq N(S).$$

Thus the region $J(S)$ determined by S is:

$$S \cup \{ac, bd, abc, bcd, cda, dab, ai\}.$$

Notice, e.g., that the covering of the cell ai by $\{ab, b, bi\}$, and hence by S, is captured by the fact that an arbitrary cell C is connected with ai only if it is connected with at least one of ab, b, bi. Also, although a necessary condition for $A \in J(S)$ is that $A \approx B$ for some $B \in S$, this condition is not sufficient. For example, $eg \notin J(S)$ since (although $eg \approx efh$) we have $fgh \approx eg$ but not $fgh \approx efh$.

We define the *sum* of regions (= *join* as in Prenowitz' system of geometry (Prenowitz and Jantosciak, 1979); cf. also Webster, 1995, for a concise account of Prenowitz' join geometry) by:

$$R + S = J(\{A \cup B | A \in R, B \in S\}).$$

This is, in effect, the region which lies between (any parts of) R and S. This operation provides us, in particular, with a simple definition of convexity:

DEFINITION 12.55 *The region R is* convex *if $R + R \subseteq R$.*

Rather than pursue the theory of convex regions, which involves some subtleties, we shall in a moment consider a notion of convex point-set, which is a little easier to develop (and to compare with existing work).

Note on terminology. We prefer not to use Prenowitz' term "join" for the operation $+$, since we do also need the join operation in the lattice of regions (and this operation is emphatically not the same as $+$). On the other hand there is a case for calling it "product", especially in view of the quantale aspect of our structures (see the remarks at the end of Sec. 7.3). We finally opt for "sum", however, for greater compatibility with the standard notation of oriented matroids, which will be important later on. If one is thinking of A, B as subsets of a Euclidean space, or more generally of a vector space over an ordered field, $A + B$ resembles an average of the two sets rather than their vector sum. Better still, it may be thought of as the set of weighted sums of elements of A, B, respectively, where the weights are positive scalars with arithmetic sum 1.

We now introduce the first of two axioms concerning points. (Incidentally, it is because of the need for these axioms, or something close to them, that it is difficult to envisage an entirely "point-free" geometry.) Geometrically, the effect of this axiom is to ensure that connection of cells means connection of

their relative interiors. Note that we often present finite sets via lists; thus, "A, p" in the following denotes $A \cup \{p\}$. Moreover we have the convention that letters $A, B, \ldots$ usually denote finite sets of points, while $p, q, \ldots$ denote singletons.

G1 If $A, p \approx B, p$ then $A \approx B$ or $A, p \approx B$ or $A \approx B, p$.

Formally, the axiom asserts that an instance of the connection relation, in which a point p occurs on both sides, can hold only because a "simpler" instance, in which p occurs on at most one side, holds. A diagram illustrating the three cases (in the conclusion of the Axiom) follows (Figs. 12.11–12.13).

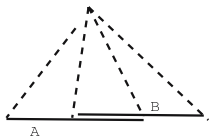


Figure 12.11. $A \approx B$.

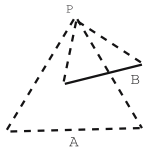


Figure 12.12. $A, p \approx B$.

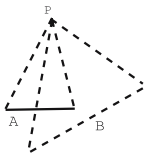


Figure 12.13. $A \approx B, p$.

The special case where B (say) is empty is allowed:

$$A, p \approx p \Rightarrow A \approx p. \tag{12.2}$$

A second axiom concerning points says, in its simplest form, that if a point is common to two regions, then the regions are connected (overlap $\Rightarrow$ connectedness). The full form of the axiom provides the general context in which the occurrence of a point may be eliminated:

G2 $p, A \approx B;\ p, A' \approx B' \ \Rightarrow\ B, A' \approx A, B'$.

The diagram (Fig. 12.14) should make it plain that this is a form of Pasch's Axiom, some form of which (as was discovered towards the end of the XIXth century) is essential for ordered geometry.

The preceding axioms on points are useful in establishing a basic property of convexity, as we shall now see.

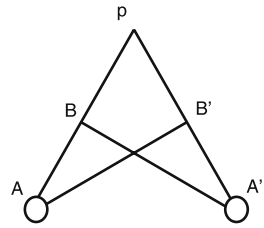


Figure 12.14. Ax. G2 (Pasch).

DEFINITION 12.56 *A set X of points is* convex *if, for any point x and cell A:*

$$x \approx A \subseteq X \Rightarrow x \in X.$$

The collection $\mathcal{C}$ of convex sets is obviously closed under intersection. Moreover the closure system involved is evidently algebraic (finitary), and the empty set is convex; so we do indeed have a convexity structure, in the sense of abstract convexity theory. The convex hull operator shall (henceforth) be denoted $[\bullet]$.

PROPOSITION 12.57 *For any set $X \subseteq G$, $[X] = \{x | \exists A \subseteq X.\ x \approx A\}$*

Proof Denote by Y the set $\{x | \exists A \subseteq X.\ x \approx A\}$. Clearly we have $X \subseteq Y \subseteq [X]$. It remains to show that Y is convex. Suppose that $p \approx s_1 s_2 \ldots s_k$, where each s_i is connected with a subset of Y. Then by Axiom G2 we can eliminate s_1 obtaining: $p \approx t_1 \ldots t_l s_2 \ldots s_k$. It is clear that we can in this way successively eliminate $s_1, \ldots, s_k$, obtaining a quasi-equation of the form $p \approx t_1 \ldots t_n \subseteq Y$. QED

We are now in a position to revisit the notions considered in the previous subsection: the faces and interior of a convex set. In the context of discrete cell-based geometry, it is natural to replace the intervals which feature in Definition 12.49 by cells. We thus arrive at:

DEFINITION 12.58 *Let $(X, \approx)$ be a cell geometry, with $C \subseteq X$ convex. Then an* interior *point of C is any point $p \in C$ such that*

$$\forall x \in C\ \exists A \subseteq C.\ p \approx A, x.$$

Further, a face *of C is any convex subset F of C such that;*

$$\forall A \subseteq C.\ A \approx F \Rightarrow A \subseteq F.$$

In order to proceed, we shall temporarily make use of the following :

PROPOSITION 12.59 *If $p \approx A (\subseteq X)$ and $q \approx B (\subseteq A)$, then $p \approx A, q$.*

This trivial-looking assertion is somewhat troublesome to prove directly from our axioms. Later it will turn out that it is an immediate consequence of an alternate axiomatization to be considered in Sec. 9 (vector axioms for oriented matroids). Hence we omit the proof for now. We have:

THEOREM 12.60 *Let $(\mathcal{P}^+_{fin}(X), \approx)$ be a cell geometry satisfying Axioms G1,G2. Let C be any convex subset of X, p a point of C, and F a subset of C. The following are equivalent:*

1 *F is the least face of C which contains p.*

2 *F is a face of C and p is an interior point of F.*

3 $F = \bigcup\{A \subseteq C \mid p \approx A\}$.

Proof Abbreviate $\bigcup\{A \subseteq C \mid p \approx A\}$ by E. Let B be any finite subset of E. Choose B' such that $p \approx B'$ and $B \subseteq B'$. Then if a is any point such that $a \approx B$, we have by the preceding proposition:

$$p \approx a, B'.$$

Thus $a \in E$, showing that E is convex.

Next, suppose that $q \in E$, and that $B \subseteq C$ is such that $q \approx B$. For some finite A' we have $p \approx q, A'$, and so by Axiom G2:

$$p \approx A, A'.$$

Thus $A' \subseteq E$, showing that E is a face. Also, it is immediate from the definition of a face that E is a subset of any face that contains p. We conclude that E is the least face containing p, that is, (1)$\Leftrightarrow$ (3).

Finally, that p is an interior point of the face F is equivalent to the assertion that every A such that $p \approx A$ is a subset of F. Hence (2) $\Leftrightarrow$ (3). QED

COROLLARY 12.61 *Every point p of a convex set C is either an interior point of C or else belongs to a proper face of C (but not both).*

7.3 Region Geometries and Oriented Matroids

In the remainder of the section, we examine some connections between our version of region geometry and two well-established areas of research. The first area we consider is that of oriented matroids. It turns out that the connection between cell geometries (developed in a certain way) and oriented matroids is very close indeed: so much so that detailed work in cell geometry is to some extent redundant, as covered by matroid theory.

To arrive at the suggested development of cell geometry (thereby bringing matroids into the picture), we may consider the algebraic structure induced by our "sum" of regions. So far we have not even shown that this sum is associative (it is evidently commutative). This will shortly be remedied. Let us for the moment consider the possible introduction of a neutral element and (quasi-)inverses into our putative semigroup of regions. It is easy enough to propose to extend the set of points with a special "zero point" e, with the stipulation:

$$A, e \approx B \;\Leftrightarrow\; A \approx B.$$

The only problem with this is that we have not, so far, allowed our quasi-equations to have one side empty, so that we are left without a meaning for:

$$e \approx A. \tag{12.3}$$

We can find models in which such formulas have a meaning by looking at spherical geometry. "Ordinary" points lie on the surface of a sphere S, while e may be thought of as the centre of S. Then (12.3) has, as usual, the meaning that e lies in the relative interior of A. Clearly, this includes the case that A consists of two antipodal points. This brings us to the interpretation of "inverses" (or involutes): we consider $-p$ to be the point antipodal to p. The rule for the inverse is that any point may be moved to the other side of a quasi-equation, with change of "sign":

$$p, A \approx B \;\Rightarrow\; A \approx B, -p.$$

Of course, we take $-(-p) = p$, and $-e = e$. It is easy to justify this rule by vector calculations, if desired (see below, Sec. 9).

The use of "signed" geometry often permits topics to be treated in a simpler and more uniform manner than is possible in "flat" geometry. In part this is a result of the following:

LEMMA 12.62 *Let S be a set of cells, and suppose that $A \in J(S)$. Then, for any cells U, V, if $U \approx A \cup V$ then $\exists W \in S. U \approx W, V$.*

Proof If $U \approx A \cup V$, then $U, -V \approx A$; since $A \in J(S)$, $U, -V \approx W$ for some $W \in S$. This gives the result. QED

Illustrating the use of this, we have:

PROPOSITION 12.63 *Sum of regions is associative.*

Proof Let R, S, T be regions, and U a cell. By definition, $U \in (R + S) + T$ iff every cell W connected with U is connected with $V \cup C$ for some $V \in$

$(R + S), C \in T$. By the preceding Lemma, this condition amounts to the assertion that every such W is connected with some cell of the form $A \cup B \cup C$, where $A \in R, B \in S, C \in T$. The analysis of "$U \in R + (S + T)$" is similar.

QED

This pattern of argument could be replicated in flat geometry only if we were, in effect, to build the above Lemma into the definition of "region". That is, we would have to define a region to be a set S of cells such that, if A is a cell such that, for any cells U, V

$$A, U \approx V \;\Rightarrow\; W, U \approx V \; for\ some\ W \in S,$$

then $A \in S$. Spherical geometry, on the other hand, lets us retain the simple view that a region is an orthoclosed subset of the connection graph.

We now come to a very significant feature of the involutory (or "spherical") calculus. The freedom we have, via the involution, to move terms to one side of a quasi-equation means that, analogously with classical sequent logic, we have the option of formulating a geometry as a calculus of *sets* of points, rather than of *pairs* of sets. The sets are called "surrounding sets", as they may be thought of as surrounding the centre e of the sphere. This will be spelled out in detail in the second part of the chapter. We write $-A$ for $\{-a|\ a \in A\}$. A quasi-equation $A \approx B$ then becomes the statement that $A, -B$ is a surrounding set. Reflexivity of $\approx$ thus states that $A, -A$ is always a surrounding set. Symmetry of $\approx$ says that whenever $A \cup -B$ is surrounding, then so is $B \cup -A$ (which is $-(A \cup -B)$). Put more simply: if C is surrounding, then so is $-C$. These are precisely Axioms S1,S2 for surrounding sets (Sec. 9). The additivity condition of Definition 12.54 translates into the statement that the union of surrounding sets is surrounding: that is, Axiom S3. Next, the cell geometry axiom G2 says that if $p, A, -B$ and $p, A', -B'$ are surrounding, then so is $B, A', -A, -B'$. If we write S, T for $p, A, -B$, $p, A', -B'$, respectively, and apply negation (inverse) to the result set, this becomes:

$$S, T\ surrounding \;\Rightarrow\; (S \setminus p) \cup (-T \setminus -p)\ surrounding.$$

That is, we have Axiom S4. We leave to the reader the task of translating cell geometry axiom G1, and checking that it corresponds to S5 of Sec. 9.

Remarks. (1) As an alternative to the "spherical" interpretation of the involutory geometry, we may choose to consider this calculus as a purely formal, conservative extension of the "flat" calculus.

(2) It is not difficult to show that the lattice of regions, taken together with the sum (now as "product"), gives us a quantale. A similar observation was made in Smyth, 2000. The setting we have now is somewhat different from that of Smyth, 2000, but it is significant that we still get a quantale of regions.

These remarks will be elaborated on another occasion.

7.4 The Ortholattice of Regions

Before concluding this brief account of region theory, it will be instructive to look at what regions amount to, concretely, in the Euclidean case. Can we identify regions with some particularly simple arrangements of cells in, say, the case of a "cell geometry" generated by a finite set of points in $\mathbb{R}^2$? In case of possible confusion, it should be emphasized that a cell geometry is by no means the same thing as a cell complex, in $\mathbb{R}^2$ or otherwise. In a cell complex the cells, as well as covering the space, must be pairwise unconnected (and also have their faces matched properly). Let us speak of a *cell partition* in the case that the cells form a pairwise orthogonal cover (regardless of whether the faces match up); see Fig. 12.15.

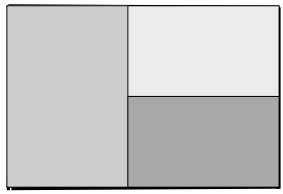


Figure 12.15. A cell partition.

One use of cell partitions is to provide us with a useful substitute for the notion that connected cells (or regions) should have points in common. Namely, we can require that, for any given cell partition Π, connected cells should be connected with a common cell of Π. Let us call this the "cell partition principle". This obviously holds in the Euclidean case. How about cell geometries in general? Notice that, in a geometry $(\Sigma, \approx)$, cell partitions can be characterized very simply: they are just maximal cliques in the orthogonality graph of Σ. (In particular, the sum plays no part.) Then we have the following characterization in terms of Dacey graphs (Definition 12.28):

PROPOSITION 12.64 *A geometry Σ satisfies the cell partition principle if and only if the orthogonality graph of Σ is Dacey.*

Proof Suppose that Π is a cell partition, that is, a maximal clique in $(\Sigma, \perp)$. Assume that $A \approx B$. Then A, B fail to be connected with a common cell of Π iff the sets $A^\perp$, $B^\perp$ form a partition of Π - which cannot happen if the Dacey condition holds. The partition principle is in effect the contrapositive of the Dacey condition. QED

Having seen the geometrical significance of the Dacey property, it is natural to ask about the stronger condition that we studied in Sec. 4.2, namely that $O\mathcal{L}(G)$ (in the present context, the lattice of regions) is orthomodular. Suppose that we have a finite (or at worst, locally finite) set N of points in Euclidean

space, which for definiteness we take as $\mathbb{R}^2$ (nothing essential depends on this restriction). We may consider the elements of the geometry, in other words the vertices of the connection graph $(\Sigma, \approx)$, to be the open cells with vertices belonging to N. As a point set, on the other hand, our space X is the convex hull of N. We seek a concrete description of the regions. Note first that any pairwise disjoint collection K of cells determines a region. More precisely, the set S of cells covered by K is a region. For suppose that B is a cell not covered by K. Let p be a point such that $p \in B - |K|$, where $|K|$ is the set $\bigcup X$ of points belonging to cells of K. Suppose without loss of generality that the set of vertices of B is a subset of $|K|$ (see Fig. 12.16).

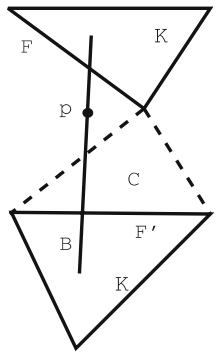


Figure 12.16. C meets the segment B but not the region K.

Then we can find two faces, say F, F', of cells in K such that the cells pF,pF' do not meet K. Using vertices of F, F', we obtain a cell (such as C in the diagram) which meets B but does not meet K. This shows that $B \notin J(K)$.

We observe that any region of the form $J(K)$ (K pairwise disjoint) has as orthocomplement a region of the same kind. For we can easily "fill up" the set $X - |K|$ with pairwise disjoint cells (in effect extending K to a cell partition of X). Indeed it is not difficult to see, via an argument similar to that in the preceding paragraph, that we can fill up the set $X - |J(K)|$ (the union of the set of cells that are disjoint from every cell of K) with pairwise disjoint cells, for every set K of cells. We conclude that any region is representable as a disjoint union of cells.

A collection of pairwise disjoint cells is, of course, the same as a clique in the orthogonality space $(\Sigma, \perp)$. Thus we conclude from the preceding theorem, together with Theorem 12.29 and Proposition 12.30 (Sec. 5.1), that the lattice of regions is orthomodular.

The conclusion we have just reached depends on the assumption that we have a geometry realized in Euclidean space. Now we have to confess that we do not know whether it holds for geometries in general. It is hoped that

further investigation will resolve this question. If it does not hold in general, we would be inclined to restrict attention to those geometries whose region lattice is orthomodular, so important and useful is this property.

To illustrate our claim about the significance of orthomodularity, we may look at triangulations. Notice that if we restrict attention to the regions which can be obtained as joins of cells belonging to a clique K of $(\Sigma, \perp)$, we obtain a Boolean algebra (isomorphic to $\mathcal{P}(K)$). Under what conditions do we have a maximal Boolean algebra of this kind? Clearly, the clique K should be maximal. But this is not enough. We must require that the cells of K cannot themselves be decomposed into disjoint unions of smaller cells. Every cell can be decomposed into simplexes; so the cells must surely be simplexes. Moreover these simplexes must be atomic: concretely, a simplex C is atomic if no points of Σ lie in any face of C (other than the vertices of C). Now a decomposition of the space into non-overlaping atomic simplexes is what is usually called a *triangulation*. At the same time, a maximal Boolean subalgebra of an orthomodular lattice is generally called a *block*. Thus we conclude that central concepts of computational geometry and of orthomodular theory are essentially the same:

$$triangulation \equiv block.$$

Unfortunately we must leave this topic here.

As we have seen, several authors have drawn attention to the difficulties involved in the attempt to develop a discrete region theory. To the previously cited authors we should add Pratt and Lemon, 1997, where it is argued that, under quite minimal conditions, a discrete theory of polygonal regions is impossible. We avoid the problems and difficulties indicated by these authors, because we abandon distributivity of the lattice of regions. But do we thereby give up too much? Can we really work with a non-distributive "logic" of regions?

Our answer is that, although the overall lattice is non-distributive, there may be, as we have seen, ample distributivity in particular contexts. Orthomodular lattices are very rich structures. Such a lattice contains, in general, many Boolean algebras. (The patching together of orthomodular lattices and posets from Boolean algebras is a major theme in orthomodular theory (Kalmbach, 1983)). Also it should not be forgotten that the regions are organized, not merely into a lattice, but into a quantale. As with any quantale, we have thereby the distributivity of the lattice join over the product.

8. Matroids

The rest of this chapter is a more conventional exposition of finite geometry, particularly finite Euclidean geometry, by which here is meant the theory of oriented matroids. The treatment of oriented matroids here is unconventional in that more attention is given to spherical rather than flat structures. The reason

for this, as also discussed in the previous section, is that finite spherical models of space are more naturally algebraic than finite flat ones. Sphericity is one of the main themes in this chapter, and possible lines of future research in the development of finite algebraic models of space will be discussed in the final section.

Oriented matroids are matroids with additional structure, so we begin with the basics of matroid theory. Of course only a tiny fraction of the theory can be discussed here, and for a far more comprehensive introduction to the subject we refer to Oxley, 1992b.

Matroids admit many axiomatizations and the following are some of the most important. Let E be a finite set:

Independent set axioms A matroid on E is a set of its subsets, called *independent sets*, satisfying:

(I1) $\emptyset$ is independent;

(I2) all subsets of an independent set are independent;

(I3) if I, J are independent and $|I| < |J|$ then there exists some $x \in J \backslash I$ such that $I \cup x$ is independent.

The *submatroid* on any $X \subseteq E$ is the set of all independent subsets of X, which clearly is again a matroid. A set is *dependent* if it is not independent. A *basis* is any maximal independent set, that is, any independent set that is not a proper subset of any other independent set. A basis of X is a basis of its submatroid. Clearly, the independent sets are precisely those sets that are subsets of bases, in terms of which the above axioms may then be translated.

Basis axioms A matroid on E is a set of its subsets, called *bases*, satisfying:

(B1) there is at least one basis;

(B2) if B, C are bases and $x \in B \backslash C$, there exists some $y \in C \backslash B$ such that $(B \backslash x) \cup y$ is a basis.

Proving that the independent set axioms imply the basis axioms is simple. Proving the converse is tricky, and involves, as with proving equivalence with some of the axioms below, hitting upon the right induction argument; see Oxley, 1992b for the full proofs of axiom equivalence.

A *circuit* is any minimal dependent set, that is, any dependent set each of whose proper subsets is independent. The independent sets are then precisely those sets that do not contain a circuit.

Circuit axioms A matroid on E is a set of its subsets, called *circuits*, satisfying:

(C1) all circuits are non-empty;

(C2) no circuit is a proper subset of any other circuit;

(C3) if $B \neq C$ are circuits and $x \in B \cap C$, then $(B \cup C) \backslash x$ contains a circuit.

It is almost immediate from the independent set axioms that any two bases of a matroid have the same cardinality, and this is quite fundamental as it means that the dimension of a matroid can be defined. The *rank* of a matroid is the cardinality of any of its bases, and the rank $r(X)$ of X is the rank of its submatroid. It is immediate that a set is independent if and only if its rank is equal to its cardinality.

Rank axioms A matroid on E is a function $r : 2^E \to \mathbb{N}$ such that:

(R1) $0 \leq r(X) \leq |X|$;

(R2) if $X \subseteq Y$ then $r(X) \leq r(Y)$;

(R3) $r(X \cup Y) + r(X \cap Y) \leq r(X) + r(Y)$.

A set X is a *flat* if it is maximal with respect to its rank, that is, if $X \subset Y$ then $r(X) < r(Y)$. It is immediate that the ground set of a matroid is a flat. Moreover, the intersection of any two flats is again a flat. This means that the *closure* $\langle X \rangle$ of a set X can be defined as the smallest flat that contains it, which is the intersection of all the flats that contain it. From the definition it is immediate that $r(\langle X \rangle) = r(X)$, and then a set I is independent if and only if $x \notin \langle I \backslash x \rangle$ for all $x \in I$.

Closure axioms A matroid on E is an operator $\langle\ \rangle$ that is increasing, monotonic and idempotent, and satisfies the *Exchange Axiom*:

(E) if $y \in \langle x \cup X \rangle \backslash \langle X \rangle$ then $x \in \langle y \cup X \rangle$.

Infinite matroids are defined simply by dropping from the closure axioms the requirement that the ground sets be finite; see Oxley, 1992a. Examples include all vector spaces, all projective spaces and all infinite undirected graphs. Moreover, the closure operator in each of these spaces is *algebraic*: the closure of a set is the union of the closures of its finite subsets. Despite the importance of these examples, the theory of infinite matroids is not as rich as that of finite matroids, which is the main subject of discussion here.

8.1 Examples

Matroids may be thought of as combinatorial geometries: *straight lines* are flats of rank two, *planes* are flats of rank three, . . . , *hyperplanes* are flats of rank one less than the rank of the whole matroid. The natural geometry in each of the following examples is a matroid geometry.

8.1.1 Matroids from vector spaces. Whitney's original paper (Whitney, 1935) introducing matroids was entitled "On the abstract properties of linear independence", and independent sets in matroids were meant to be abstract, combinatorial versions of linear independent sets in vector spaces. For basic vector space definitions and theory, see any standard text such as Mac Lane and Birkhoff, 1967.

Let E be any finite subset of any vector space V, and let the independent sets be those subsets of E that are linear independent in V. That this is a matroid, called the matroid of *linear dependences* on E, follows from basic vector space dimension theory. It may be checked that the flats in the matroid are the intersections with E of the subspaces of V, that the rank of any X is its linear dimension, and that $\langle X \rangle$ is the intersection with E of the vector subspace $span(X)$. An example is shown in Fig. 12.17.

The construction can be generalized to any finite *multiset* E of elements of V by letting the independent sets be those subsets of E not containing repeated elements that are linear independent. Multisets are likely to be encountered when working with *matrices*, from which the word *matroid* was coined.

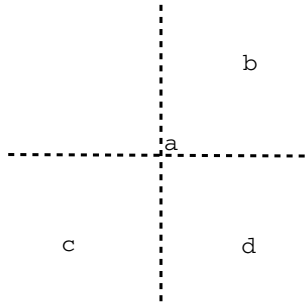


Figure 12.17. The linear independent sets in $abcd$ are b, c, d, bd, cd.

The construction does not necessarily give a matroid in more general modules over rings (rather than over fields). In a module over a ring, it can happen that two maximal independent sets have different cardinalities, making dimension difficult to work with. This cannot happen in a vector space, and one of the main advantages of vector spaces over modules over rings is that they have a better theory of dimension.

A second matroid, of *affine dependences*, on E is obtained by taking the independent sets to be those of its subsets that are affine independent in V. This construction is not essentially different to the above, however, as a set $x_1, \ldots, x_k$ is affine dependent in $V(= F^n)$ if and only if $(x_1, 1), \ldots, (x_k, 1)$ is linear dependent in the vector space F^{n+1}.

(The affine independent sets in Fig. 12.17 are: $a, b, c, d, ab, ac, ad, bc, bd, cd, abd, acd, bcd$.)

8.1.2 Matroids from graphs. Graphs have a natural geometry. In any undirected graph, which may have loops and multiple edges, a set of edges is a *circuit* if it is non-empty and can be ordered $(e_0, \ldots, e_k)$ such that there exist vertices $v_0, \ldots, v_k$, all distinct, such that each e_i is incident to v_i and $v_{(i+1)}$ (indices considered $mod\ k$). By this definition, all loops ($k = 0$) are circuits. The matroid derived from the graph has as its ground set the set of edges, and its circuits are the circuits in the graph. It is a simple exercise to show that this is a matroid. An example is shown in Fig. 12.18.

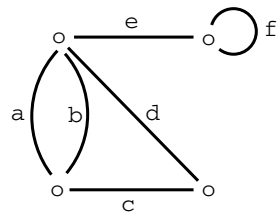


Figure 12.18. An undirected graph, with circuits ab, acd, bcd, f.

A set of edges is a *forest* of if it contains no circuit. A forest is *spanning* if every vertex is incident to at least one of its edges. The forests are then the independent sets of the graph matroid, and the spanning forests are the bases. With regard to duality (to be discussed shortly), the cocircuits of the graph matroid are the minimal cut-sets of the graph.

8.2 Projective spaces

A *projective space* is a set together with a collection of its subsets, called *lines*, such that:

(P1) any two distinct points are contained in exactly one line;

(P2) there is more than one line;

(P3) for any distinct points p, q, r, s, t, if there is a line containing p, q, r and a line containing p, s, t, then the line containing q, s intersects the line containing r, t.

A *subspace* is any set that for each two of its points contains the straight line passing through them. The set of subspaces is the set of flats of a matroid (the infinite case included). Moreover, the projective spaces are then precisely those matroids whose rank function is *modular*: $r(X \cup Y) + r(X \cap Y) = r(X) + r(Y)$. (The rank axiom $(R3)$ is the requirement that the rank function in a matroid be *sub-modular*.) For the details, and a full-blown treatment of projective geometry in terms of matroids, see Faure and Frölicher, 2000.

8.3 Matroid duality

Arguably the single most important and fundamental aspect of matroid theory is that every matroid has another, dual matroid on its ground set.

THEOREM 12.65 *The collection of complements of the bases of a matroid is itself the collection of bases of a matroid.*

Proof The basis exchange axiom (B2) implies (and is in fact equivalent to):

(B2*) if B, C are bases of a matroid and $x \in C \backslash B$ then there is $y \in B \backslash C$ such that $(B \backslash y) \cup x$ is a basis.

This is *not* obtained from the exchange axiom simply by relabelling, but by: $B \cup x$ is dependent so contains a circuit X, x is independent so X contains some $y \in B$, and then $(B \backslash y) \cup x$ does not contain a circuit, so is independent.

Now let B, C be bases and $x \in (E \backslash B) \backslash (E \backslash C) = C \backslash B$. There is some $y \in B \backslash C = (E \backslash C) \backslash (E \backslash B)$ such that $(B \backslash y) \cup x$ is a basis. Finally, $E \backslash ((B \backslash y) \cup x) = ((E \backslash B) \backslash x) \cup y$. QED

Complements of bases are called *cobases*, and the matroid they give is called the *dual matroid* of the original matroid. Obviously, every matroid is the dual of its dual. Circuits, flats, etc., of the dual matroid are called *cocircuits*, *coflats*, etc., of the original matroid Matroid duality is a spatial property that is revealed at the finite level. Oxley, 1992b remarks that "It [matroid duality] massively increases the weapons at one's disposal in attacking any matroid problem. The theory of infinite matroids provides ample evidence of the vital role that duality plays in matroid theory. In the infinite context, there is no duality theory having the richness of the theory in the finite context."

The following result is an important geometric aspect of matroid duality, the significance of which becomes more apparent in the context of oriented matroids.

PROPOSITION 12.66 *The cocircuits of a matroid are the complements of its hyperplanes.*

Proof A subset of a matroid is *spanning* if its closure is the whole ground set, that is, if it contains a basis. Then a set is spanning if and only if its complement is coindependent. Then: H is a hyperplane if and only if H is not spanning but, for all $x \notin H$, $H \cup x$ is spanning, if and only if $E \backslash H$ is coindependent but, for all $x \in E \backslash H$, $(E \backslash H) \backslash x$ is codependent. QED

The following exercises in basic matroid theory conclude this brief introduction:

(1) *Uniform matroids.* On a set of cardinality n, for any $m \leq n$ let the independent sets be those subsets of cardinality $\leq m$. Verify that this is a matroid. Find the bases, circuits and flats of this matroid. Show that the dual of a uniform matroid is again uniform.

(2) *Partition matroids.* With respect to any equivalence relation on a finite set, say that a set is independent if no two of its elements are equivalent. Verify that this is a matroid. Find the bases, circuits and flats of this matroid. Show that the cocircuits are the equivalence classes.

(3) *Flats and closure.* Let I be independent and let $A = \{x \mid I \cup x \text{ is dependent}\}$. Using only the independent set axioms, prove that:

(a) $r(I \cup A) = r(I)$;

(b) $I \cup A = \langle I \rangle$

(c) for any set X, $\langle X \rangle = \{x \mid r(X) = r(X \cup x)\}$;

(d) the intersection of two flats is again a flat.

(4) *Matchings.* A *matching* in an undirected graph is any set of edges maximal with respect to the property that no two edges have a common vertex. Let $\{X, Y\}$ be a bipartition of a bipartite graph. Say that a set of edges is X*-independent* if no two of its edges have a common vertex in X, and is Y*-independent* if no two have a common vertex in Y. Show that:

(a) the X-independent sets are the independent sets of a matroid, as are the Y-independent sets;

(b) a matching is any set of edges maximimal with respect to being both X- and Y-independent;

(c) the collection of sets that are both X- and Y-independent is not in general the collection of independent sets of a matroid.

(5) *The lattice of flats is graded.* Prove that the intersection of two flats is again a flat, and so that the collection of flats ordered by subset inclusion is a lattice. Prove that this lattice is *graded*, that is, that all its maximal flags (chains) are of the same length. (Hint: consider sequences $I_1 \subset \cdots \subset I_k$ of independent sets.)

(6) *Matroids and diagram geometry.* Let $F \subset G$ be flats of a matroid such that $r(G) = r(F) + 3$. Show that for any distinct flats X_1, X_2 such that $F \subset X_1, X_2 \subset G$ and $r(X_1) = r(X_2) = r(F) + 1$, there exists a unique flat Y such that $X_1, X_2 \subset Y \subset G$. A *linear geometry* is a collection of points and

straight lines that satisfies Axioms (P1), (P2) of projective geometry. Conclude that every interval of length 3 in the lattice of flats of a matroid is a linear geometry.

(7) *Coproducts of matroids.* Let E_1, E_2 be ground sets of matroids and let E denote their disjoint union. Define a subset of E to be independent if it is the union of an independent set in E_1 with an independent set in E_2. Show that this gives a matroid.

(8) *Graphic matroids are vectorial.* Given an undirected graph with m vertices and n edges, consider the $(m \times n)$-matrix whose rows and columns are labelled by vertices and edges respectively, and whose (i, j)-th entry is 1 if the ith row label is incident to the jth column label, and is 0 otherwise. Consider the columns as points of the vector space $(Z_2)^m$. Show that the resulting matroid of linear dependences is the circuit matroid of the graph.

9. Spherical oriented matroids

Matroids are geometries of linear or affine sets (straight lines, planes, ...), but do not have any structure of convexity. There is no notion of *betweenness* on the points of a straight line or, dually, hyperplanes do not separate the space into two *half-spaces*. Matroids therefore have to be given additional structure, an *orientatation*, from which convexity may be derived, and this is the theory of oriented matroids.

Convexity is not anyway to be expected in matroids, as general vector spaces do not have convexity. It is only vector spaces over ordered fields, such as the field of real numbers, that do. Each straight line in a vector space is isomorphic to the field, and betweenness on the line is derived from the linear order on the field.

There are two conceptions of an oriented matroid, *flat* and *spherical*, the definitions of both appearing more or less independently in the same journal in the same year. Flat oriented matroids, due to Bland and Vergnas, 1978, are far more commonly studied than spherical oriented matroids, which are due to Folkman and Lawrence, 1978. Spherical oriented matroids are relegated to exactly one exercise in the standard, classic textbook on the subject Björner et al., 1993. This is justified by the fact that the two concepts are equivalent, though some care has to be taken over expressing this equivalence formally.

9.1 Involutions

It seems that the concept of an *involution* is basic in combinatorial spherical geometry. One thinks of the involute of a point on the surface of a sphere as its

opposite, or *antipode*. In Euclidean geometry, the involute of any such point x is simply its negation $-x$.

Abstractly, an involution on a set is any endomorphism such that, for all x, $x = -(-x)$ and $x \neq -x$. The latter *fixed-point-free* requirement that $x \neq -x$ is not usually assumed, but there is no reason not to assume it here. A set together with an involution will be called an *involuted set*. For any subset X of an involuted set, its involute $-X$ is the set $\{-x \mid x \in X\}$. X is *admissible* if $X \cap -X = \emptyset$, and *symmetric* if $X = -X$. A *transversal* is a maximal admissible set, that is, any set that for all x contains exactly one of $x, -x$. Clearly, every finite involuted set has an even cardinality.

The natural geometric interpretation of a finite set of cardinality n is as the set of vertices of the n-simplex, which may be thought of as the convex hull in $\mathbb{R}^n$ of its canonical basis, which is the set of points $e_1, \ldots, e_n$, where e_i is the point that has ith coordinate 1 and all other coordinates 0. Subsets of the set then correspond exactly to the sets of vertices of the faces of the n-simplex: the face lattice of the n-simplex is the Boolean algebra over n elements.

On the other hand, the natural geometric interpretation of an involuted set of cardinality $2n$ is as the set of vertices of the n-*crosspolytope*, which is the convex hull of the set of points $e_1, \ldots, e_n, -e_1, \ldots, -e_n$. The admissible sets then correspond exactly to the sets of vertices of the faces of the n-crosspolytope (see Fig. 12.19). Thus the lattice of admissible sets (together with an adjoined top element) is the face lattice of the crosspolytope, which is the dual (i.e. the lattice turned upside-down) of the face lattice of the n-cube.

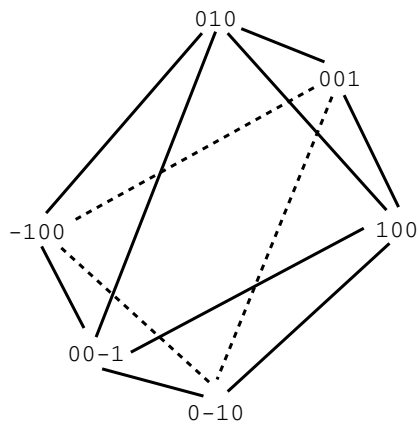


Figure 12.19. The 3-crosspolytope (octahedron).

9.2 Spherical geometry

To re-cap the basics of classical spherical geometry, let S_n denote the n-dimensional unit sphere, that is, the set of all points in $\mathbb{R}^{n+1}$ at distance one from the origin. The natural involution on the sphere is negation.

A *ray from the origin* is any set of the form $\{\lambda x \mid \lambda \geq 0\}$, for any $x \neq 0$. There is an obvious bijection between the set of all such rays and the sphere. A *convex cone* is any convex set that is the union of such rays. A subset of the sphere is *spherical convex* if it is the intersection of a convex cone with the sphere, and *spherical linear* if it is the intersection of a subspace with the sphere. An attractive feature of spherical geometry is that linearity may be derived from convexity together with the involution, which is thematic in what is to follow.

THEOREM 12.67 *A set is spherical linear if and only if it is spherical convex and symmetric with respect to the involution.*

Real projective space P_n may be derived from S_n by quotienting. Consider any transversal (taken with respect to the involution) of the sphere, and say that its subspaces are its intersections with the spherical linear sets. This results in P_n, no matter which transversal is chosen.

A spherical linear set is a *spherical hyperplane* if it is the intersection of the sphere with a hyperplane that contains the origin (i.e. an n-dimensional subspace). The two corresponding *open hemispheres* are the respective intersections of the hyperplane's open half-spaces with the sphere. A *closed hemisphere* is the union of an open hemisphere with its spherical hyperplane ("at infinity"). Any open hemisphere together with its inherited geometry is isomorphic to $\mathbb{R}^n$.

Spherical geometry has been adopted by Stolfi, 1991, as a geometry that has convexity while at the same time retaining many of the advantages of projective geometry, which does not have convexity. The main advantage is the retention of (a version of) homogeneous coordinates.

Convexity in Euclidean geometry is algebraic: the convex hull of a set is the union of the convex hulls of its finite subsets. The same is true in spherical geometry. The *spherical relative interior* of a finite subset $X = \{x_1, \ldots, x_k\}$ of the sphere is the set of all $x \in S_n$ such that there exist scalars $\lambda_i > 0$ such that $x = \sum \lambda_i x_i$. The *spherical convex hull* of X is the set of all $x \in S_n$ such that there exist scalars $\lambda_i \geq 0$, such that $x = \sum \lambda_i x_i$.

THEOREM 12.68 *A set is spherical convex if and only if it contains the spherical convex hull of each of its finite subsets.*

The set X is *surrounding* if the origin is in its relative interior, that is, if there exist scalars $\lambda_i > 0$ such that $0 = \sum \lambda_i x_i$. It is simple to prove that x is in the

spherical relative interior of X if and only if $X \cup -x$ is surrounding, and that the spherical convex hull of X is the union of the respective spherical relative interiors of all subsets of X.

9.3 Spherical oriented matroids

It has just been shown that all of the convex and linear geometry of the sphere can be derived from the surrounding sets together with the involution. Oriented matroid theory may be regarded as an abstract, combinatorial theory of this. It is convenient to now adopt the convention that the empty set is surrounding. Say that a *vector* is any admissible surrounding set, and that a *circuit* is any minimal non-empty vector. These are the intuitions for the following definitions, the first of which is due to Folkman and Lawrence, 1978.

Circuit axioms A *spherical oriented matroid* is a finite involuted set together with a collection of its subsets, called *circuits*, such that:

(C1) each circuit is non-empty and admissible;

(C2) no circuit is a proper subset of any other;

(C3) if C is a circuit then so is $-C$;

(C4) if $C \neq D$ are circuits and $x \in C \cap D$ then $(C \backslash x) \cup (-D \backslash -x)$ contains a circuit.

A *vector* in an oriented matroid is any set that is both admissible and the union of circuits. The empty union is allowed, and so $\emptyset$ is always a vector. Vector axioms for oriented matroids come in various apparently differing strengths; the following is a version of the Vector Axioms (4.1.1) in Björner et al., 1993.

Vector axioms A spherical oriented matroid is a finite involuted set together with a collection of its subsets, called *vectors*, such that:

(V1) $\emptyset$ is a vector;

(V2) each vector is admissible;

(V3) if U is a vector then so is $-U$;

(V4) the union of vectors, if admissible, is again a vector;

(V5) if U, V are vectors and $x \in U \cap V$ then $(U \backslash x) \cup (-V \backslash -x)$ contains a vector that contains $(U \backslash V) \cup (-V \backslash -U)$.

A *surrounding set* in an oriented matroid is any set that is the union of a vector and a symmetric set. The following axiomatization, referred to in Sec. 7 above, appears to be new.

Surrounding set axioms A spherical oriented matroid is a finite involuted set together with a collection of its subsets, called *surrounding sets*, such that:

(S1) every symmetric set is surrounding;

(S2) if S is surrounding then so is $-S$;

(S3) the union of surrounding sets is again surrounding;

(S4) if S, T are surrounding and $x \in S \cap T$ then $(S \backslash x) \cup (-T \backslash -x)$ is also surrounding;

(S5) if S is surrounding and $x, -x \in S$ then at least one of the sets $S \backslash x$, $S \backslash -x$, $S \backslash \{x, -x\}$ is also surrounding.

9.4 Equivalence of the axiomatizations

Axiom (C4) is called the *Circuit Elimination Condition*. The following apparently stronger condition is a most important basic property of oriented matroids, though as the proof of the result runs to well over a page in both Folkman and Lawrence, 1978 and Björner et al., 1993; it will not be repeated here. Using only the Circuit Axioms:

LEMMA 12.69 *If $C \neq D$ are circuits and $x \in C \cap D$ and $y \in (C \backslash D) \cup (-D \backslash -C)$ then $(C \backslash x) \cup (-D \backslash -x)$ contains a circuit that contains y.*

The *composition* of a pair of subsets X, Y of an involuted set is the set $X \circ Y = X \cup (Y \backslash -X)$. This operation is associative but not commutative. The composition of two admissible sets is again admissible, and composition might be thought of as a kind of non-symmetric union on admissible sets. Again, using only the Circuit Axioms:

LEMMA 12.70 *The composition of two vectors is again a vector.*

Proof The result amounts to the fact that if U, V are vectors and $x \in V \backslash (U \cup -U)$ then $x \in C \subseteq U \circ V$ for some circuit C. Let C be any circuit satisfying (a) $x \in C$ and (b) $C \subseteq V \cup U \cup -U$, and is such that $|C \cap -U|$ is minimal among all such circuits that satisfy. If $C \cap -U$ is empty then the result is proved, so suppose for contradiction that $y \in C \cap -U$. Let $D \subseteq -U$ be a circuit that contains y. Then $x \in C \backslash D$ and, by strong elimination, the set $(C \backslash y) \cup (-D \backslash -y)$ contains a circuit that satisfies (a), (b) and contradicts minimality of $|C \cap -U|$. QED

PROPOSITION 12.71 *The Circuit Axioms are equivalent to the Vector Axioms.*

Proof Defining vectors in terms of circuits it is easy to see that (V1)–(V4) are satisfied. For (V5), let U, V be vectors and $x \in U \cap V$. If $U = V$ then $\emptyset$ satisfies (V5). If not, it is first claimed that, for any $y \in U \backslash V$, there is a circuit C_y that contains y and satisfies (a) $C_y \cap (-U \backslash -V) = \emptyset$ and (b) $C_y \cap (V \backslash U) = \emptyset$. To prove this, if there is a circuit $C \subseteq U$ that contains y but not x, let $C_y = C$. If not, let $D \subseteq V$ be any circuit that contains x. Then $y \in C \backslash D$ so, by strong elimination, there is a circuit $C_y \subseteq (C \backslash x) \cup (-D \backslash -x)$ with $y \in C_y$, and this proves the claim. Now taking the composition (in any order) of all the circuits C_y results in a vector W such that $U \backslash V \subseteq W$ and $W \cap (-U \backslash -V) = \emptyset$. By a similar construction on $-V \backslash -U$, there is a vector W' such that $-V \backslash -U \subseteq W'$ and $W' \cap (V \backslash U) = \emptyset$. The compostion $W \circ W'$ then satisfies the conditions of (V5).

Conversely, beginning with the Vector Axioms and defining circuits as minimal non-empty vectors, it is easy to see that the Circuit Axioms are satisfied. It remains to show that every vector is the union of circuits, which will be proved by induction on vector size. Let U be a vector and assume that the property holds for all vectors of cardinality $< |U|$. It is first claimed that, for any $x \in U$, there is a circuit $C \subseteq U \cup -U$ such that $x \in C$. Let $D \subseteq U$ be any circuit and assume that $x \notin D$ (otherwise the claim is proved). For any $y \in D$, by vector elimination there is a vector $V \subseteq (U \backslash y) \cup (-D \backslash -y)$ that contains x. Then $|V| < |U|$ because vectors are admissible, so, by the induction hypothesis, the claim is proved. Now let C be any circuit such that (a) $x \in C$ and (b) $C \subseteq U \cup -U$, and is such that $|C \cap -U|$ is minimal among all such circuits. Suppose for contradiction that $y \in C \cap -U$. By vector elimination there is a vector $V \subseteq (C \backslash y) \cup (U \backslash -y)$ that contains x. By the induction hypothesis, V contains a circuit D that contains x. Then D satisfies (a), (b), but $|D \cap -U|$ contradicts minimality of $|C \cap -U|$. QED

PROPOSITION 12.72 *The Vector Axioms are equivalent to the Surrounding Set Axioms.*

Proof Beginning with the Vector Axioms and defining surrounding sets in terms of vectors, it is easy to show that (S1)–(S3) are satisfied. For (S4), let U, V be vectors and $x \in U \cap V$. Let W be a vector such that $(U \backslash V) \cup (V \backslash U) \subseteq W \subseteq (U \backslash x) \cup (-V \backslash -x)$. The union of W with the symmetric set $((X \cup -X) \cap (Y \cup -Y)) \backslash \{x, -x\}$ is the set $(U \backslash x) \cup (-V \backslash -x)$. If U is a vector, X a symmetric set and $x \in U \cap X$, then $(U \backslash x) \cup (-X \backslash -x) = U \cup X \backslash \{x, -x\}$. If X, Y are symmetric sets and $x \in X \cap Y$ then $(X \backslash x) \cup (-Y \backslash -x) = X \cup Y$. Proving (S4) is then a matter of checking cases. For (S5), let U be a vector and X a symmetric let, and let $x, -x \in U \cup X$. If $x, -x \notin U$ then $(U \cup X) \backslash \{x, -x\} = U \cup X \backslash \{x, -x\}$. If $x \in U$ then $(U \cup X) \backslash -x = U \cup X \backslash \{x, -x\}$. If $-x \in U$ then $(U \cup X) \backslash x = U \cup X \backslash \{x, -x\}$.

Conversely, beginning with the Surrounding Set Axioms and defining a vector as an admissible surrounding set, it is easy to see that (V1)–(V4) are satisfied. For (V5), if U, V are vectors and $x \in U \cap V$ then $(U \backslash x) \cup (-V \backslash -x)$ is surrounding, and a vector satisfying (V5) is obtained by repeated application of (S5). Repeated application of (S5) may also be used to show that every surrounding set may be recovered as the union of a vector and a symmetric set. QED

We can now supply a straightforward proof of Proposition 12.59, whose proof was omitted in Sec. 7. In our current notation (and slightly generalized) the proposition may be rewritten:

PROPOSITION 12.73 *Let A be an admissible set, B a subset of A, and suppose that $A \cup -p$ and $B \cup -q$ are surrounding sets. Then $A \cup \{q, -p\}$ is surrounding.*

Proof Note first that if $q \in B$ there is nothing to prove. Likewise if $-q \in B$ the result is immediate, since in that case $A \cup \{q, -p\} = A \cup -p \cup \{q, -q\}$, and thus is surrounding. Assume then that neither q nor $-q$ occurs in B. Similarly we may assume that $p \neq q$.

There are now two (main) cases. Suppose first that p does not occur in A. Then $A \cup -p$ is (admissible hence) a vector, and (in virtue of (V3)) $q \cup -B$ is a vector. Hence the composition $(A \cup -p) \circ (q \cup -B) = A \cup \{q, -p\}$ is a vector. In the second case, suppose that $p \in A$. Then by (S5) at least one of $A, A \setminus p, (A \cup -p) \setminus p$ is surrounding (hence a vector). Consider the compositions $A \circ (q \cup -B)$, $(A \setminus p) \circ (q \cup -B)$, $((A \cup -p) \setminus p) \circ (q \cup -B)$. In each case $A \cup \{q, -p\}$ results from taking the union of $\{p, -p\}$ with the composition, and hence is surrounding. QED

9.5 Examples

9.5.1 Realizable oriented matroids. It is left as an exercise to show that the circuits on the sphere satisfy (C1)–(C4), that the vectors on the sphere satisfy (V1)–(V5) and that the surrounding sets on the sphere satisfy (S1)–(S5). It is then immediate that, for any finite subset E of the sphere that is symmetric with respect to the involution, the collections of circuits, vectors and independent sets that are subsets of E satisfy the respective oriented matroid axioms. The oriented matroids that can be constructed in this way are called *realizable*. An example is given in Fig. 12.20.

9.5.2 Hyperplane arrangements. A second construction of realizable oriented matroids is given by *hyperplane arrangements*. A hyperplane arrangement in $\mathbb{R}^{n+1}$ is a finite set of hyperplanes, each of which contains the origin. Consider the set E of the corresponding open half-spaces, with the obvious

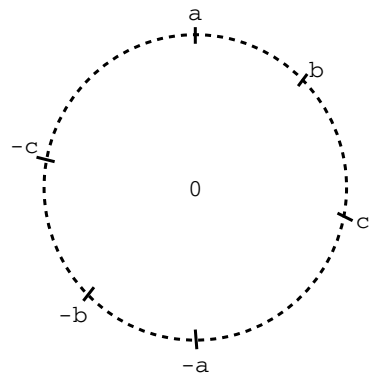


Figure 12.20. The circuits are $a - bc$ and $-ab - c$.

involution. For any point $x \in \mathbb{R}^{n+1}$, consider the set of all the half-spaces that contain x. The set of all subsets of E that are constructed in this way is the set of vectors of an oriented matroid. An example is given in Fig. 12.21.

This example has a natural spherical interpretation. Each hyperplane is considered as a spherical hyperplane, and the half-spaces are then considered as spherical open hemispheres. Taking two points on the sphere to be equivalent if they induce the same oriented matroid vector (i.e. if they both lie in exactly the same set of hemispheres), the set of equivalence classes is a cellular decomposition of the sphere.

The hyperplane arrangement construction is in fact dual to the construction of the oriented matroid on a symmetric point set E on the sphere. For each $x \in E$, consider the the open half-space containing x of the hyperplane orthogonal to x. Labelling this open half-space x and then taking the hyperplane arrangement construction gives the dual (see oriented matroid duality, to follow) oriented matroid of the oriented matroid on E.

9.6 Convexity and linearity

Convexity in oriented matroids is derived from surrounding sets just as in spherical geometry.

DEFINITION 12.74 *For any subset X of an oriented matroid:*

(a) its relative interior *is the set $relint(X)$ of all points x such that $X \cup -x$ is surrounding;*

(b) its convex hull *is the union $[X]$ of the respective relative interiors of all subsets of X.*

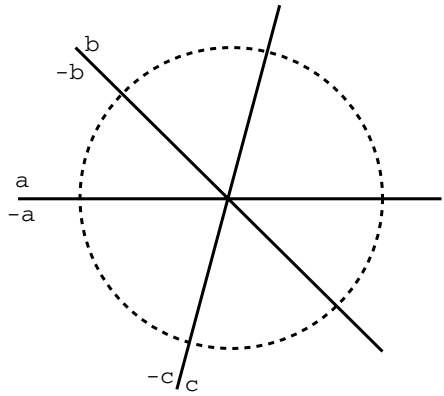


Figure 12.21. With the half-spaces labelled in this way, the vectors of the oriented matroid of this hyperplane arrangement are $\emptyset, abc, bc, -abc, -ac, -a-bc, -a-b, -a-b-c, -b-c, a-b-c, a-c, ab-c, ab$.

Thus $x \in [X]$ precisely when there exists a surrounding set S such that $-x \in S \subseteq -x \cup X$.

PROPOSITION 12.75 *The convex hull operator is increasing, monotonic and idempotent.*

Proof It is clear that, for any point x, $x \in relint(x)$, and so the convex hull operator is increasing. (It is *not* in general true that X and $relint(X)$ are comparable.) It is immediate from the definition that the convex hull operator is monotonic. For idempotency, let $x \in [[X]]$. There is some $Y = \{y_1, \ldots, y_k\} \subseteq [X]$ such that $-x \cup Y$ is surrounding. For each i there is some $X_i \subseteq X$ such that $-y_i \cup X_i$ is surrounding. By elimination on y_1, the set $-x \cup Y \backslash y_1 \cup X_1$ is surrounding. Eliminating then on $y_2, \ldots, y_k$ it follows that $-x \cup X_1 \cup \cdots \cup X_k$ is surrounding, and so $x \in [X]$. QED

DEFINITION 12.76 *A subset of an oriented matroid is* convex *if it is equal to its convex hull, and is* linear *if it is both convex and symmetric.*

As the convex hull operator is that of a closure system, the intersection of convex sets is again convex. As also the intersection of symmetric sets is again symmetric, the intersection of linear sets is again linear. Clearly, the ground set is linear, so the *linear hull* $\langle X \rangle$ may be defined as the smallest linear set that contains X.

The proofs of the next two results are left as a straightforward exercise in the manipulation of surrounding sets.

PROPOSITION 12.77 *For any set X and any x, y:*

(a) $-relint(X) = relint(-X)$*;*

(b) $-[X] = [-X]$

(c) $[X \cup -X] = \langle X \rangle$*;*

(d) $[x \cup X] \cap [-x \cup X] = [X]$*;*

(e) $y \in [x \cup X] \Rightarrow -x \in [-y \cup X]$

PROPOSITION 12.78 *For any linear* L *and any* $x, y \notin L$*:*

(a) $\langle L \cup x \rangle = [L \cup x] \cup [L \cup -x]$*;*

(b) $y \in [L \cup x] \Leftrightarrow x \in [L \cup y]$ *(*$\Leftrightarrow [L \cup x] = [L \cup y]$*);*

(c) $-x \notin [L \cup x]$.

PROPOSITION 12.79 *Surrounding sets have the following properties:*

(a) a set is surrounding if and only if its convex hull is linear;

(b) if S *is surrounding and* C *is convex then* $S \subseteq C$ *implies* $-S \subseteq C$.

Proof (a) If S is surrounding then, for any $x \in -S$, $-x \in S \subseteq -x \cup S$, and so $x \in [S]$. Then $-S \subseteq [S]$, and then $\langle S \rangle = [S \cup -S] = [S]$. For the converse, if $[X] = \langle X \rangle$ then, for all $x \in X$, $-x \in [X]$, which means that there is some surrounding set S such that $x \in S \subseteq x \cup X = X$. Then X is the union of surrounding sets and so is itself surrounding.

(b) Because $-S \subseteq \langle S \rangle = [S] \subseteq C$. QED

PROPOSITION 12.80 *The linear sets are the flats of a matroid.*

Proof That the linear hull operator is the closure operator of a matroid is a simple consequence of convexity property (e). QED

This matroid is called the *linear matroid* of the oriented matroid. Each set $\{x, -x\}$ is dependent, therefore every independent set is admissible. A consequence of the fact that every linear set is symmetric is that every reorientation (see later) of an independent set is again independent. Then every reorientation of a matroid circuit is again a matroid circuit. The oriented matroid circuits do not in general satisfy the circuit axioms of a matroid, but the circuits of the linear matroid are precisely the reorientations of the oriented matroid circuits, together with those sets $\{x, -x\}$ such that x is not dependent.

9.7 Oriented matroid duality

As with matroids, duality is a most important part of oriented matroid theory.

DEFINITION 12.81 *Subsets X, Y of an involuted set are* orthogonal*, written $X \perp Y$, if: $X \cap Y = \emptyset \Leftrightarrow X \cap -Y = \emptyset$.*

The orthogonality relation is symmetric, a set is admissible if and only if it is not orthogonal to itself, and a set is symmetric if and only if it is orthogonal to all sets.

A subset of an oriented matroid is *cosurrounding* if it is orthogonal to every surrounding set. *Covectors* are admissible cosurrounding sets, and *cocircuits* are minimal non-empty covectors.

PROPOSITION 12.82 *The set of cosurrounding sets of an oriented matroid satisfies the oriented matroid surrounding set axioms.*

Proof A set is symmetric iff it is orthogonal to all sets, which gives (S1). If $X \perp Y$ then $X \perp -Y$, which gives (S2). If $X \perp Y, Z$ then $X \perp (Y \cup Z)$, which gives (S3).

For (S4), suppose for contradiction that X, Y are cosurrounding, $x \in X \cap Y$ and $(X \backslash x) \cup (-Y \backslash -x)$ is not cosurrounding. Then, without loss of generality, there exists a surrounding set S such that (a) $(X \backslash x) \cap S \neq \emptyset$, (b) $(X \backslash x) \cap -S = \emptyset$ and (c) $(Y \backslash x) \cap S = \emptyset$. That $X \perp S$ gives that (d) $X \cap -S = x$, which together with (c) gives that (e) $Y \cap S = x$. So $x, -x \in S$, and then from (b) it follows that $-x \notin X$. Then from (a) it follows that there is $y \neq x, -x$ such that $y \in X \cap S$. But Y and $S \backslash x$ are not orthogonal because $Y \cap (S \backslash x) = \emptyset$ and $x \in Y \cap (-S \backslash -x)$; X and $S \backslash -x$ are not orthogonal because $X \cap (-S \backslash x) = \emptyset$ and $x \in X \cap (S \backslash -x)$; X and $S \backslash \{x, -x\})$ are not orthogonal because $X \cap (-S \backslash \{x, -x\}) = \emptyset$ and $y \in X \cap (S \backslash \{x, -x\})$. This contradicts the property (S5) of surrounding sets.

For (S5), let X be cosurrounding, let $x, -x \in X$ and suppose for contradiction that there exist surrounding sets R, S, T such that: (a) $(X \backslash x) \cap R \neq \emptyset$ and $(X \backslash x) \cap -R = \emptyset$; (b) $(X \backslash -x) \cap S \neq \emptyset$ and $(X \backslash -x) \cap -S = \emptyset$; (c) $(X \backslash \{x, -x\}) \cap T \neq \emptyset$ and $(X \backslash \{x, -x\}) \cap -T = \emptyset$. From (a) and $X \perp R$ it follows that (d) $X \cap -R = x$. From (b) and $X \perp S$ it follows that (e) $X \cap -S = -x$. Then $x \in -R \cap S$, so $(-R \backslash x) \cup (-S \backslash -x)$ is surrounding. X is disjoint from this set, and so is disjoint from its involute. It follows that (f) $X \cap R = -x$, and (g) $X \cap S = x$. From (c) and $X \perp T$ it may be assumed without loss of generality that $-x \in -T$. Then $x \in S \cap T$, so $(-S \backslash -x) \cup (T \backslash x)$ is surrounding. By (c), X intersects $T \backslash x$, and therefore intersects the involute $(S \backslash x) \cup (-T \backslash -x)$. By (g), X therefore intersects $(-T \backslash -x)$, and then, by (c), x is the only possible point in the intersection. Therefore both $x, -x \in T$. By (c), X and $T \backslash \{x, -x\}$ are not orthogonal. If X and $T \backslash x$ were orthogonal then,

by (f), $-x \in R \cap (T \backslash x)$. Then $(-R \backslash x) \cup (T \backslash \{x, -x\})$ would be surrounding, but, by (c), X intersects this set, and, by (f) and (c), does not intersect its involute. Finally, if X and $T \backslash -x$ were orthogonal then, by (g), $x \in S \cap (T \backslash -x)$. Then $(-S \backslash -x) \cup (T \backslash \{x, -x\})$ would be surrounding, but, by (c), X intersects this set, and, by (g) and (c), does not intersect its complement. Therefore none of $T \backslash x$, $T \backslash -x$, $T \backslash \{x, -x\}$ are surrounding, a contradiction. QED

PROPOSITION 12.83 *Cosurrounding sets have the following properties:*

(a) complements of convex sets are cosurrounding;

(b) complements of admissible cosurrounding sets (i.e. covectors) are convex.

Proof (a) Follows immediately from property (b) of surrounding sets. (b) Straightforward exercise. QED

The oriented matroid of cosurrounding sets is called the *dual* of the original oriented matroid.

THEOREM 12.84 *Every oriented matroid is the dual of its dual.*

Proof Let X be orthogonal to all cosurrounding sets. The complement $E \backslash [X]$ of the convex set $[X]$ is cosurrounding and disjoint from X, therefore $X \cap -(E \backslash [X]) = \emptyset$. Now $-(E \backslash [X]) = E \backslash - [X] = E \backslash [-X]$, and so $X \subseteq [-X]$. Then $[X]$ is linear, so X is surrounding. QED

9.7.1 Example: Oriented matroids from Euclidean subspaces. The orthogonality relation on a finite involuted set has the following geometric interpretation in terms of Euclidean space considered with its standard inner product. Recall that two points $x, y \in \mathbb{R}^n$ are orthogonal, written $x \perp y$, if their inner product is zero. Consider now the involuted set as the set of vertices of the crosspolytope, as discussed earlier, and, for any of its subsets X, let X_c denote the corresponding set of crosspolytope vertices. It is straightforward though tedious to prove that $X \perp Y$ if and only if there exist $x \in relint(X_c)$ and $y \in relint(Y_c)$ such that $x \perp y$. It is left as an exercise in classical Euclidean geometry to prove that:

PROPOSITION 12.85 *For any subspace S of $\mathbb{R}^n$, the set of all faces of the crosspolytope whose relative interior intersects S is the set of vectors of an oriented matroid. Moreover, the covectors are those faces whose relative interior intersects the orthogonal subspace $S^\perp$.*

9.8 Hemispheres

The structure of a dual oriented matroid is bound-up with the following concept.

DEFINITION 12.86 *For any hyperplane H in the linear matroid of an oriented matroid, its* closed hemispheres *are the sets $[H \cup x]$ and $[H \cup -x]$, for any $x \notin H$. Its* open hemispheres *are the complements of its closed hemispheres.*

According to the convexity properties in the previous section, for the definition of a closed hemisphere it makes no difference which $x \notin H$ is chosen. It is a simple consequence of the convexity and linearity properties that the union of the two closed hemispheres is the whole ground set, the intersection of the two closed hemispheres is the hyperplane, one closed hemisphere is the involute of the other, and closed hemispheres are convex but not linear. To summarize:

PROPOSITION 12.87 *Where H^-, H^+ are the open hemispheres of the hyperplane H:*

(a) $H^+ = -(H^-)$;

(b) $\{H^-, H, H^+\}$ is a partition of the ground set;

(c) the two closed hemispheres are $H \cup H^-, H \cup H^+$;

(d) $H \cup H^- = [H \cup x]$ for any $x \in H^-$.

PROPOSITION 12.88 *The open hemispheres of an oriented matroid are precisely its cocircuits.*

Proof If a closed hemisphere $H \cup H^-$ contains a surrounding set S then, by property (b) of surrounding sets, so does its involute $H \cup H^+$. Then $S \subseteq H$, and therefore every open hemisphere is cosurrounding. Conversely, if U is admissible and cosurrounding then $L = E \backslash (U \cup -U)$ is linear because it is both symmetric and the intersection of two convex sets. If U is non-empty then $L \neq E$, and so L is contained in some hyperplane H. The remainder of the proof relies on the following result from Folkman and Lawrence, 1978. QED

LEMMA 12.89 *If C, D are circuits of an oriented matroid such that $C \subseteq D \cup -D$ then $C = D$ or $C = -D$.*

Proof Suppose for contradiction that C intersects both D and $-D$ and then show that C must then properly contain another circuit.) To complete the proof, the open hemisphere H^- is a subset of $U \cup -U$. Then, when U is a cocircuit, one of $U, -U$ is equal to H^-, and then the other is equal to H^+. QED

The following is a combinatorial version of Weyl's Theorem.

PROPOSITION 12.90 *The convex sets in an oriented matroid are precisely the intersections of its closed hemispheres.*

Proof The ground set is the intersection of the empty collection of closed hemispheres. The complement of a convex set C is cosurrounding. Then if $x \notin C$ there is a cocircuit that contains x and is disjoint from C. Cocircuits are open hemispheres, and the complement of an open hemisphere is a closed hemisphere. QED

9.9 Polytopes

If $\mathbb{R}^n$ is used as the basic mathematical model in spatial computation then polytopes are quite fundamental in that they are regions that can be finitely specified (i.e. as convex hulls of finite sets). For a treatment of polytopes and oriented matroids, see Ziegler, 1995. Within a finite spatial model, it is tempting to say that the polytopes are simply the convex sets. This is certainly true if the model is flat, though more care must be taken with spherical models.

DEFINITION 12.91 *A* polytope *in an oriented matroid is any set that is both convex and admissible.*

The requirement of admissibility constrains polytopes to be subsets of closed hemispheres, and so no non-empty subspace is a polytope. Without this constraint, it would not be possible to obtain a version of, for example, the Krein-Milman theorem (see below).

Say that a *face* of a polytope P is any of its subsets $F = P \backslash U$, where U is cosurrounding and $P \cap -U = \emptyset$. Consider, for example, the case where U is a cocircuit (i.e. an open hemisphere). Then, where H is the corresponding hyperplane, P is contained in the closed hemisphere $H \cup U$ and $F = P \cap H$. Some simple properties of faces are:

PROPOSITION 12.92 *Let P be a polytope:*

(a) both $\emptyset$ and P are faces of P;

(b) all faces of P are again polytopes;

(c) the intersection of two faces of P is again a face of P.

Proof (a) $E \backslash P$ and $\emptyset$ are both cosurrounding. (b) Any face is the intersection of an admissible convex set with a linear set $E \backslash (U \cup -U)$. (c) Straightforward exercise. QED

Combinatorial versions of several classical convexity theorems, including Weyl's Theorem above, have been proved for flat oriented matroids; see Björner et al., 1993. A version of Caratheodory's theorem in the spherical case can be obtained for admissible sets using the fact that if X is admissible and $x \notin X$, then $x \in [X]$ precisely when there is a circuit C such that $-x \in C \subseteq X \cup -x$, and then relating rank to circuits as for matroids. For the Krein-Milman theorem, say that an *extreme point* of a polytope P is any $x \in P$ such that $P \backslash x$ is convex. It has been shown for flat oriented matroids that any convex set is the convex hull of its extreme points, and a spherical version can be obtained via the "equivalence" of flat and spherical oriented matroids (to be discussed).

9.9.1 Anti-matroids. A second important concept in combinatorial convexity is that of an *anti-matroid*, which is a set together with a closure (monotonic, increasing and idempotent) operator [ ] that satisfies the *Anti-Exchange property*

(AE) for all $x \neq y$ and all X, if $x \in [y \cup X] \backslash [X]$ then $y \notin [x \cup X]$.

This should be compared with the exchange axiom for matroids. The main models are any subsets of Euclidean space with the inherited convex hull operator. The Anti-Exchange property has been considered as a foundation for digital geometry by Pfaltz and others (Pfaltz, 1996; see also Coppel, 1998).

The anti-exchange property does not hold in spherical geometry; for a counter-example, let X be a closed hemi-sphere and let x, y be distinct points in the opposite open hemi-sphere: then $[x \cup X] = [y \cup X] =$ the whole sphere. The property does however hold locally on the sphere, as the induced geometry on an open hemi-sphere is Euclidean. Spherical oriented matroids do not either have the anti-exchange property, though they do have the following local version (see also Las Vergnas, 1980).

PROPOSITION 12.93 *Let X be a subset and x, y points of an oriented matroid that are all contained in some cosurrounding set, and let $[x] \neq [y]$. If $y \in [x \cup X] \backslash [X]$ then $x \notin [y \cup X]$.*

Proof If $y \in [x \cup X] \backslash [X]$ then $y \notin [X]$, so there is some $S \subseteq X$ such that $-x \cup y \cup S$ is surrounding. Suppose for contradiction that $y \in [x \cup X]$, that is, that there is some $T \subseteq X$ such that $-y \cup x \cup T$ is surrounding. By elimination on x, the set $y \cup -y \cup S \cup T$ is surrounding. But $-y \cup S \cup T$ is not surrounding because $y \notin [X]$, and $y \cup S \cup T$ is not surrounding because it is a non-empty subset of a cosurrounding set. Therefore $S \cup T$ is surrounding, and so, by orthogonality, must be empty. But then $\{-x, y\}$ is surrounding, and then $x \in [y]$, a contradiction. QED

10. Flat oriented matroids

The "flat" oriented matroids of Bland & Las Vergnas are far more commonly studied than spherical oriented matroids. The two concepts are equivalent, though some care has to be taken over expressing this equivalence formally.

Flat oriented matroids do not require any involution, but the price for this is the use of signed sets. A *signed subset* of a set is an ordered pair (X^+, X^-) of its subsets such that X^+, X^- are disjoint. Its *support* is the set $X^+ \cup X^-$ and its *opposite* is the signed set (X^-, X^+).

Oriented matroid axioms A flat oriented matroid is a finite set together with a collection of its signed subsets, called *signed circuits*, satisying:

(OM1) the opposite of a signed circuit is again a signed circuit;

(OM2) the set of supports of the signed circuits is the set of circuits of a matroid;

(OM3) if $(X^+, X^-) \neq (Y^+, Y^-)$ are signed circuits and $x \in X^+ \cap Y^-$ then there exists a signed circuit (Z^+, Z^-) such that $Z^+ \subseteq (X^+ \cup Y^+)\backslash x$ and $Z^- \subseteq (X^- \cup Y^-)\backslash x$.

The underlying matroid of a flat oriented matroid has as its circuits the set of supports of the signed circuits. The oriented matroid is said to be an *orientation* of its underlying matroid. Not all matroids can be oriented; in particular, *no* projective space can be oriented.

For an understanding of the definition, consider that one means of adding convexity to a matroid is to partition the complement of each of its hyperplanes into two *open half-spaces*, with a *closed half-space* then being the union of an open half-space with its hyperplane. A *convex set* is then defined as any set that is the intersection of a collection of closed half-spaces. Recall that the complements of hyperplanes in a matroid are precisely its cocircuits, so a definition of open-half spaces involves splitting each cocircuit in two. When the cocircuits are considered as circuits of the dual matroid, this results eventually in the axioms above.

10.1 Examples

10.1.1 Realizable flat oriented matroids. Consider the matroid of linear dependences on a finite subset E of Euclidean space. Any circuit $\{x_1, \ldots, x_k\}$ of this matroid may be converted into a pair of opposite signed circuits as follows. For any linear dependence $\lambda_1 x_1 + \cdots + \lambda_k x_k = 0$, let $X^+ = \{x_i \mid \lambda_i > 0\}$ and $X^- = \{x_i \mid \lambda_i < 0\}$. The collection of all signed circuits constructed in this way is an orientation of the matroid of linear dependences. An example is given in Fig. 12.22.

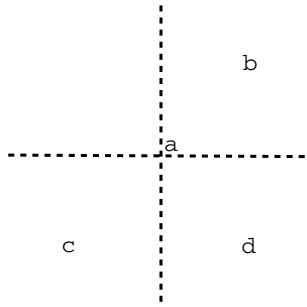


Figure 12.22. The signed circuits are $(\emptyset, a), (a, \emptyset), (b, c), (c, b)$.

Now consider the matroid of affine dependences on E. Any circuit of this matroid may be converted into a pair of opposite signed circuits as follows. A *Radon partition* of the circuit (or indeed any other set) is any partition of it into two sets whose respective relative interiors intersect. Each circuit admits a unique Radon partition, which is then converted into a pair of opposite signed circuits. The collection of all signed circuits constructed in this way is an orientation of the matroid of affine dependences. (The signed circuits for the point set in Fig. 12.22 are $(a, bc), (bc, a)$.)

10.1.2 Flat oriented matroids from graphs. Consider a directed graph, which may have loops and multiple edges. A set of edges is a circuit if it is a circuit in the undirected graph obtained by forgetting edge orientation. Moving around a circuit, from one edge to the next, say that an edge is *positive* if it is oriented in the direction one is moving, and *negative* if it is oriented against this direction. This gives a partition of the circuit, which is then considered as a pair of opposite signed circuits. The collection of all signed circuits constructed in this way is an orientation of the graph matroid. An example is given in Fig. 12.23.

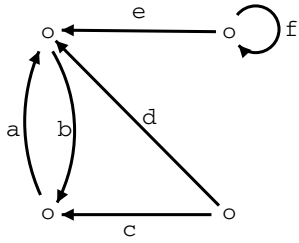


Figure 12.23. The signed circuits are $(ab, \emptyset), (\emptyset, ab), (ac, d), (d, ac), (bd, c), (c, bd)$, $(f, \emptyset), (\emptyset, f)$.

10.2 Reorientation

A morphism between involuted sets E, F is any function $f : E \to F$ such that, for all x, $f(-x) = -f(x)$. A *symmetry* of E is any of its bijective endomorphisms. Geometrically, this is symmetry of the crosspolytope.

Any symmetry on an involuted set E extends in the obvious way to a bijection $2^E \to 2^E$. For any such bijection there exists a symmetric set $S \subseteq E$ such that the bijection is of the form $X \mapsto (X \backslash S) \cup (-X \cap S)$. Conversely, any symmetric set gives such a bijection.

DEFINITION 12.94 *A* reorientation *of a spherical oriented matroid is the image of all its surrounding sets under a symmetry.*

It is a simple exercise that the reorientation of a spherical oriented matroid is again a spherical oriented matroid.

Reorientation helps explain the relationship between spherical and flat oriented matroids. A flat oriented matroid may be constructed from a spherical one as follows. Let T be any transversal, and say that the signed circuits in T are those signed sets of the form $(C \cap T, -C \cap T)$, for all circuits C. This gives a flat oriented matroid, the underlying matroid of which is the sub-matroid on T with respect to the underlying matroid of the spherical oriented matroid. This construction is analogous to the construction of real projective space from the sphere discussed earlier.

Conversely, beginning with a flat oriented matroid on E, make a copy $-E$ and say that a circuit in $E \cup -E$ is any set of the form $X^+ \cup -X^-$, for all signed circuits (X^+, X^-). This gives a spherical oriented matroid.

Despite the simplicity of these translations, some care has to be taken over the equivalence of the two concepts. To see this, consider the important notion of a flat oriented matroid being *acyclic*. A signed set (X^+, X^-) is said to be *positive* if $X^- = \emptyset$, and a flat oriented matroid is acyclic if it contains no positive signed circuit. Every realizable flat oriented matroid, for example, is acyclic. It cannot make any sense, however, to say that a spherical oriented matroid is acyclic. Whether or not an acylic flat oriented matroid is derived from a spherical oriented matroid depends entirely on the choice of the transversal (in fact, there is always some transversal that is acyclic).

The *reorientation* of a signed set (X^+, X^-) at a set S is the signed set $((X^+ \backslash S) \cup (X^- \cap S), (X^- \backslash S) \cup (X^+ \cap S))$. Beginning with a flat oriented matroid and any subset of the ground set, the reorientation of all signed circuits at this set gives a set of signed circuits of another oriented matroid. The relation of one flat oriented matroid being a reorientation of another is an equivalence relation. Beginning with a spherical oriented matroid, the collection of all flat oriented matroids derived from all transversals is an equivalence class. Conversely, any two spherical oriented matroids equivalent under reorientation

have the same equivalence class of flat oriented matroids. Therefore spherical and flat oriented matroids are the same concept *up to reorientation*.

10.3 Oriented matroids and computational geometry

One application of flat oriented matroids is as a foundation for computational geometry. This is the approach taken by Knuth, 1991, who argued that working axiomatically in this way gives better algorithm design. Knuth worked with the following definition.

CC-System Axioms A CC-system is a finite set together with a ternary relation (write xyz to mean that (x, y, z) is in the relation) such that:

(CC1) $xyz \Rightarrow yzx$,

(CC2) $xyz \Rightarrow \neg xzy$,

(CC3) $xyz \vee yxz$,

(CC4) $zxy \wedge wzy \wedge wxz \Rightarrow wxy$,

(CC5) $zyv \wedge zyw \wedge zyx \wedge zvw \wedge zwx \Rightarrow zvx$.

The main intended models are as follows. Consider any finite subset of the plane that does not contain any three collinear points. For any three points x, y, z, consider the unique circle that contains these points on its circumference. Then let xyz if, when moving around the circumference in a clockwise direction, after hitting x one hits y before z. An example is given in Fig. 12.24.

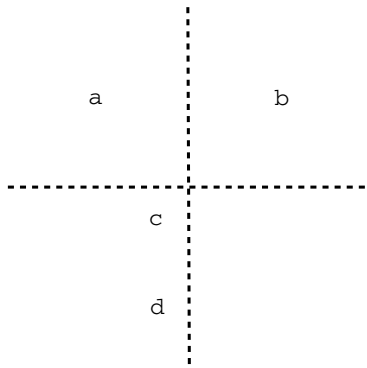


Figure 12.24. The ternary clockwise relation is abc, bca, cab, abd, bda, dab, acd, dac, cad, bdc, dcb, cbd.

CC-systems have the following flat oriented matroid structure. A *hyperplane* is any set that contains exactly two points. The complement of a hyperplane

$\{x, y\}$ is called a *cocircuit*, and may be partitioned into the two sets $\{z \mid xyz\}$, $\{z \mid xzy\}$. The resulting set of signed cocircuits satisfies the flat oriented matroid axioms. In fact, CC-systems are equivalent to rank 3 flat oriented matroids in which all sets of cardinality 3 are bases of the underlying matroid.

11. Algebraic spatial models

It seems by now to be fairly well understood what the basic mathematical structures involved in discrete topology are: graphs, finite (or locally finite) posets (T_0-spaces) and simplicial or cell-complexes. These, anyway, are those structures employed routinely for representation of Euclidean topology in computational areas of application such as geographical information systems, spatial database theory, digital topology and image analysis.

It has been argued here that these structures themselves may usefully be considered as instances of closure spaces of various types, and also that this approach facilitates unification with classical continuous topology.

In extending topology to geometry, it seems that by far the most sophisticated finite models of the linear and convex structure of Euclidean space are oriented matroids. Finite geometry is a broad area of research, but much of it is not Euclidean in flavour and no theory seems to match that of oriented matroids in establishing combinatorial Euclidean geometry.

Oriented matroid theory has some serious deficiencies, however. Arguably the most serious is that there is no general notion of morphism, and consequently no category. Even if there were, it is difficult to see how even basic constructions such as products could be obtained. Oriented matroids also have a quite severe lack of point-extensions, which is a serious impediment to any theory of refinement.

The spherical version of oriented matroids has been presented here because it might be more conducive to a development of finite algebraic models of Euclidean geometry. The question is how far can the combinatorial method in Euclidean spatial representation be pushed? Must it stop at convexity and linearity? Must, for example, metrical structure always be expressed using $\mathbb{R}$? Given that the research programmes of developing finite Euclidean geometry and topology have been largely successful, a logical progression would seem to be to try to express this finite structure algebraically. This would parallel modern classical geometry in placing algebra (vector spaces) first, topology and geometry being derived from the algebra.

One would hope that an algebraic treatment might lead to a reasonable category of finite spatial models, the present lack of which has been stressed several times in this chapter. Any such category would be expected to at least have products, as surely any true version of Euclidean geometry must ultimately involve the construction of higher-dimensional spaces as products of lower-dimensional

ones. Classically, this is only possible algebraically. Any purely synthetic approach to Euclidean geometry does not, it seems, have this construction. It is only taking products of vector spaces that works. The combinatorial analogue of this is that matroids do not, apparently, have products.

A more directly practical motivation is that algebraic spatial representation might facilitate better storage and computation of spatial objects. For example, it is probably fair to say that Knuth's oriented matroid approach to computational geometry has not yet caught on. Knuth himself regards oriented matroids as a useful tool for computing the geometry of finite point sets in $\mathbb{R}^n$, and points in an oriented matroid are (at least when doing computational geometry) to be regarded as points in $\mathbb{R}^n$. The ternary clockwise relation is computed using standard determinants of matrices computation on such points, and collinear triples of points are explicitly ruled out of the axiom system precisely because betweenness of points in $\mathbb{R}^n$ that may have irrational coordinates is not decidable in finite time. The interesting point here is the idea of a *background*: as in so much of spatial computation, $\mathbb{R}^n$ is the mathematical background in terms of which spatial objects and locations are ultimately considered. Oriented matroids, though they are alternative finite backgrounds, are in general far too large to feasibly be stored. It is far simpler to store a point set as a set of coordinates in $\mathbb{R}^n$ and compute their geometry algebraically (e.g. by calculating determinants of matrices), rather than store explicitly a basic set of their spatial relations (the oriented matroid). It might be that finite coordinate systems for computational geometry, if such structures can exist, would retain the undoubted merits of Knuth's approach while avoiding explicit storage of too much information.

Our specification for a finite algebraic Euclidean geometry is any well-understood finite algebra that has a natural oriented matroid structure. There is no shortage of these in 1-dimension. Consider any ring $\mathbb{Z}_{2n}$: a natural involution is $x \mapsto x + n$, and it is fairly clear that the oriented matroid of the ring considered as a point set on the circle may be derived from its algebra.

The graph pictured is a natural cyclic order (on a cyclic group). The isomorphisms of this graph are the rotations $x \mapsto m + x$, and the anti-isomorphisms (i.e. order-reversing) are the reflections $x \mapsto m - x$. The rotations together with the reflections form of course a (Coxeter) group, namely the *Dihedral group* of order $2n$. $\mathbb{Z}_{2n}$ is then simply the subgoup of rotations.

The aim now is to extend this to higher dimensions. Recall that convexity in Euclidean geometry is derived algebraically from the fact that the field $\mathbb{R}$ is ordered. Spherical orders have convexity too, one that is locally (i.e. on open hemispheres) Euclidean, and this is part of the theory of spherical oriented matroids. One might therefore hope that the modules $(\mathbb{Z}_{2n})^k$ have spherical convexity, to be derived from the cyclic structure of the ring of scalars $\mathbb{Z}_{2n}$ in a way analogous to the construction of classical Euclidean convexity out of the

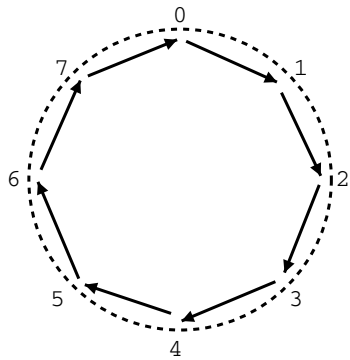


Figure 12.25. $\mathbb{Z}_8$ considered as points on a circle (think of a clockface).

linear order on $\mathbb{R}$. The geometry of these modules (as of finite vector spaces) is, however, not spherical but toroidal: the product of two circles is not a sphere but a torus. We do not see, moreover, how a spherical geometry may naturally be constructed out of a toroidal one.

One the other hand, the product of two *discs* (filled-in circles) *is* spherical. (In fact it is a simple result that the product of two Euclidean convex sets is again convex.) Might there be some way of regarding $\mathbb{Z}_{2n}$ as a disc rather than a circle and deriving the spherical structure of products accordingly? Consider the simplest case, that of $\mathbb{Z}_2$ considered as an interval (a 1-dimensional disc). The product of k intervals is the k-cube, whose vertices certainly sit on the surface of the $(k+1)$-sphere and therefore have an oriented matroid structure. But this structure seems to resist all attempts to understand it! It is astounding that one of the most regular, symmetric objects (a platonic solid) in mathematics has a natural Euclidean geometry that in higher dimensions can only, it seems, be computed by brute force. Aichholzer and Aurenhammer, 1996, for example, estimate that the running time for determining the number of subsets of vertices of the 9-cube that generate hyperplanes in $\mathbb{R}^9$ is about 35 years.

Problem Find a combinatorial description of the geometry of the n-cube.

Any such description would surely have some bearing on the open:

Las Vergnas Cube Conjecture The linear matroid on the vertices of the n-cube has, up to reorientation, exactly one orientation.

What is doubly annoying here is that any *hypergraph* (i.e. a finite set together with any collection of its subsets) has a natural Euclidean geometry, a spherical oriented matroid structure, as the vertices of the n-cube correspond naturally

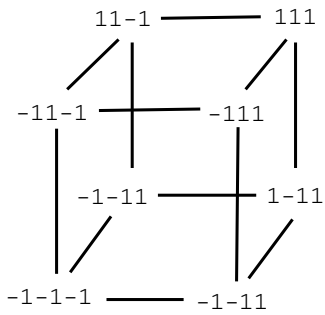


Figure 12.26. The 3-cube, whose spherical oriented matroid circuits are $\{-1-1-1, 11-1, -111, 1-11\}$ and $\{111, -1-11, 1-1-1, -11-1\}$. Higher dimensional cubes have many more circuits, which seem to resist combinatorial characterization.

to subsets of a set of cardinality n. Whether or not this geometry is of interest or has meaning in any context one cannot really say, since there is no existing good description of it.

To compound the issue, the spherical oriented matroid on the vertices of the dual of the n-cube, namely the n-crosspolytope, is utterly trivial: it is the oriented matroid in which only the symmetric sets are surrounding. (The dual of this oriented matroid is that in which all sets are surrounding.)

11.1 Symmetry

One would expect any spherical formulation of space to have a natural symmetry, akin to that it inherits when considered as embedded on the Euclidean sphere. If this formulation were algebraic, one might hope that the symmetry group and the algebra would be closely related. The problem is, of course, that the symmetry group of the Euclidean n-sphere has, in dimensions higher than one, essentially very few n-dimensional finite subgroups, certainly not enough to re-generate the group. This is an apparent violation of the (generalized) Correspondence Principle, and this lack of finite symmetry is perhaps the real challenge to the finitist programme of combinatorial spatial representation.

A way forward, however, might emerge out of a recent approach to matroids in A. V. Borovik and White, 2003 that is wholly algebraic and based entirely on symmetry. What follows is a brief presentation of this.

11.2 The Gale order

For any finite set E of cardinality n and for any $0 \leq m \leq n$, let E_m denote the set of all subsets of cardinality m.

For any linear order $\leq$ on E and any $X, Y \in E_m$, whose respective inherited orderings are $x_1 < \cdots < x_m$ and $y_1 < \cdots < y_m$, let $X \leq_m Y$ if, for each i,

$x_i \leq y_i$. The order $\leq_m$ is clearly a partial order, and is called the *Gale order* on E_m determined by $\leq$.

It is easy to prove that the Gale order is always a lattice ordering. Matroids can be defined in terms of the Gale order; the following result is in A. V. Borovik and White, 2003.

THEOREM 12.95 *Let $\mathcal{B}$ be a non-empty subset of E_m. Then $\mathcal{B}$ is the collection of bases of a matroid if and only if for each linear order on E the partial order on $\mathcal{B}$ induced by the Gale order has a top element.*

Proof *(Only if:)* Assume for contradiction that for some linear order there exist $B \neq C \in \mathcal{B}$ that are both maximal in the induced Gale order. By the basis exchange axiom, for any $x \in B \backslash C$ there is some $y \in C \backslash B$ such that $B \backslash x \cup y \in \mathcal{B}$. Then $x > y$ otherwise B would not be maximal. By a similar argument, for any $y \in C \backslash B$ there is some $x \in B \backslash C$ such that $y > x$, clearly giving a contradiction.

(If:) Let $B, C \in \mathcal{B}$ and let $x \in B \backslash C$. If $B \backslash C = x$ then $C \backslash B = y$ and $(B \backslash x) \cup y = C$. If not, then B, C are incomparable in the Gale order of the linear order $x < \cdots < c_1 < \cdots < c_j < b_1 < \cdots < b_k$, where the c_i's are the elements of $C \backslash B$ and the b_i's are the elements of $B \backslash x$. There is some element of $\mathcal{B}$ above both, and this can only be $\{b_1, \ldots, b_k\} \cup$ some c_i. QED

11.3 The Greedy algorithm

Condider a matroid together with a linear order on its ground set. The order typically is induced by a mapping of the ground set into the reals; for example, a cost function. The following is an efficient means of finding the maximum basis in the induced Gale order on the set of bases:

Greedy algorithm

```
Input: A matroid with a linear order on its ground set
  I := ∅
    loop
      Find the largest x ∉ I such that I ∪ x is independent
      I := I ∪ x
      Stop when there is no x ∉ I such that I ∪ x is independent
    end loop
  Return I
```

THEOREM 12.96 *The Greedy algorithm always returns the maximum basis in the induced Gale order on the set of bases.*

Proof Every independent set is contained in a basis, so the algorithm returns a basis. Let the return be $\{x_1, \ldots, x_m\}$, where $x_1 > \cdots > x_m$. As all subsets of a basis are independent, each x_i was chosen on the ith loop. Suppose for contradiction that there is a basis $y_1 > \cdots > y_m$ that is not below the returned basis in the Gale order. Let i be the least index such that $y_i > x_i$. Now $\{y_1, \ldots, y_i\}$ and $\{x_1, \ldots, x_{i-1}\}$ are both independent, so there is some $y_j \notin \{x_1, \ldots, x_{i-1}\}$ such that $\{x_1, \ldots, x_{i-1}, y_j\}$ is independent. But $y_j \geq y_i > x_i$, contradicting that x_i was chosen on the ith loop. QED

The Greedy algorithm is greedy because it always chooses the largest element available on each loop. It is efficient because this is never a bad choice. Matroids can be defined in terms of the Greedy algorithm; the definition is similar to that in terms of the Gale order.

11.4 The Symmetric groups

A linear order on a set may be interpreted as a permutation (a.k.a. bijection, isomorphism) of its elements, and there is a one-to-one correspondence between the set of all linear orders on a finite set (say, of cardinality n) and the set of its permutations. To see this, choose any linear order $x_1 < \cdots < x_n$ and associate this with the identity function. Any other linear order $y_1 < \cdots < y_n$ is then associated with the permutation $x_i \mapsto y_i$.

The characterization of a matroid in terms of the Gale order involves quantification over *all* linear orders on a set. The set of linear orders, interpreted as the set of permutations, forms a group, with the group operation being function composition. This is the *symmetric group* on n elements, Sym_n, which is the isomorphism group of a set of cardinality n. The theory of Coxeter matroids develops matroid theory in terms of the symmetric groups alone, the Gale order being in this context the *Bruhat order*, which has a natural algebraic and geometric meaning. The following is a brief account of the Bruhat order, without proofs; a full account may be found in A. V. Borovik and White, 2003, Brown, 1989, Ronan, 1989.

Consider Sym_n explicitly as the isomorphism group of the set $[n] = \{1, \ldots, n\}$. The group operation is denoted $\circ$. For $1 \leq i < n$, let s_i denote the isomorphism that swaps $i, i+1$ and leaves all other elements fixed. The set $S = \{s_1, \ldots, s_{n-1}\}$ then generates Sym_n. The corresponding *Cayley graph* has as vertices the elements of Sym_n, and there is an edge between vertices u, v if there is some generator s_i such that $v = u \circ s_i$. This graph has no loops, and is undirected since $v = u \circ s_i \Rightarrow u\ (= u \circ s_i \circ s_i) = v \circ s_i$. There is at most one i such that $v = u \circ s_i$, and so the edge between u, v may be labelled unambiguously by i. Each vertex therefore has degree $n-1$. Another name for this graph is the *n-permutahedron*. The graph is highly symmetric, as each left group action is a graph isomorphism: for any w, there is an (i-)edge between

u, v if and only if there is an (i-)edge between $w \circ u, w \circ v$. An example is shown in Fig. 12.27.

This all has the following geometric interpretation. Sym_n is the symmetry group of the n-simplex, the elements of $[n]$ being interpreted as the vertices of the n-simplex. The elements of Sym_n may themselves be interpreted as simplices on the surface of the n-simplex, namely the simplices obtained by taking the barycentric subdivision of each facet of the n-simplex. Each such simplex is an $(n-1)$-simplex, and defining any two such simplices to be adjacent if they have a common $(n-2)$-face gives the n-permutahedron.

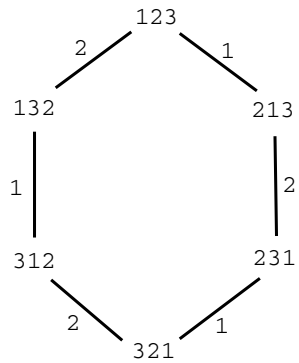


Figure 12.27. The Cayley graph of Sym_3, obtained by associating the linear order 123 with the identity.

11.4.1 The Bruhat order. A *reflection* in Sym_n is any conjugate of any element of the generating set, that is, any element of the form $w \circ s \circ w^{-1}$, where $w \in Sym_n$ and $s \in S$. Concretely, the reflections are the isomorphisms that swap i, j for some $i \neq j$ and leave all other elements fixed. Geometrically, the reflections are the reflections of the n-simplex.

The *wall* of a reflection r is the set of all edges (u, v) in the Cayley graph such that $r \circ u = v$ (equivalently $r \circ v = u$, since $r \circ r = id$). Each edge is contained in exactly one wall. Each wall cuts the Cayley graph into exactly two components, called the *hemispheres* (or *roots*) of the reflection. Concretely, where r is the reflection that swaps i, j, one of its hemispheres is the set of all linear orders in which $i < j$ and the other is the set of all linear orders in which $j < i$. Each reflection maps one hemisphere bijectively onto the other. Geometrically, where elements of Sym_n are considered as simplices in the barycentric subdivision of the n-simplex, a half-space of a reflection is the set of all simplices on the same side of the reflecting hyperplane. An example is shown in Fig. 12.28.

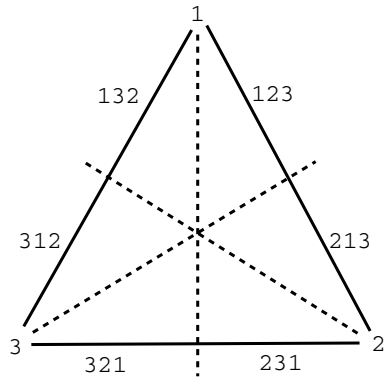


Figure 12.28. The triangle, together with planes of reflection and induced barycentric subdivision of facets.

A *folding* is "half" a reflection: one half-space is fixed while the other is reflected onto it. Formally, where r is a reflection with half spaces H^-, H^+, the corresponding *folding* of H^+ onto H^- is the map $Sym_n \to Sym_n, x \mapsto x$ if $x \in H^-, x \mapsto r \circ x$ if $x \in H^+$. Each reflection has precisely two corresponding foldings. Every folding is a graph morphism.

DEFINITION 12.97 *For any element w, the* Bruhat order *on Sym_n with bottom element w is defined as: $u \leq_w v$ if there exist foldings $f_1, \ldots, f_k$, each towards a half-space that contains w, such that $u = f_1 \circ \cdots \circ f_k(v)$.*

The Bruhat order is clearly a partial order. For any $K \subseteq \{1, \ldots, n-1\}$, a *$K$-residue* is any component of the graph obtained from the Cayley graph by removing all edges not labelled by some $i \in K$. Equivalently, a K-residue is, in Sym_n, any left coset of the subgroup generated by the set $\{s_i \mid i \in K\}$. It happens that, for any w, any residue has both a top and bottom element in its inherited Bruhat order. Moreover, for any K-residues R, S and any Bruhat order $\leq_w$, the following are equivalent: the bottom element of $R \leq_w$ the bottom element of S; the top element of $R \leq_w$ the top element of S; some element of $R \leq_w$ some element of S. Setting $R \leq_w S$ if these properties hold gives a partial order, called the Bruhat order for w on the set of K-residues.

DEFINITION 12.98 *A* Coxeter matroid *for a symmetric group is a collection $\mathcal{B}$ of K-residues such that, for any $w \in Sym_n$, $\mathcal{B}$ has a top element in its partial order induced by the Bruhat order $\leq_w$.*

This is in fact a generalization of the definition of a matroid. Matroids arise from the special case that $K = \{1, \ldots, n-1\} \backslash i$. In this case, there is a bijection between the K-residues and the subsets of $\{1, \ldots, n\}$ of cardinality i. The Bruhat order and the Gale order then coincide.

In fact, Coxeter matroids are more general still. The above definitions, of Cayley graph, reflection, wall, half-space, folding, Bruhat order, etc., work for any Coxeter group.

The following exercise might help with some of the above:

In the Boolean algebra over n elements, say that two maximal flags (a.k.a. maximal chains) are adjacent if they differ at exactly one element. Show that this gives the n-permutahedron. Say, moreover, that two adjacent maximal flags are i-adjacent if the element at which they differ has height i. Show that this is i-adjacency in the permutahedron.

Acknowledgments

Thanks are due to the other members of our Digital Geometry group, namely Giovanni Curi, Xiang Feng, Iain Stewart and Rueiher Tsaur, for discussions of the material of this chapter, and for much help in its preparation.

The second author would particularly like to thank Prof. Michel Las Vergnas and all members of l'Equipe Combinatoire at Paris 7 for their kind and generous support during his stay in Paris.

References

A. V. Borovik, I. M. Gelfand and White, N. (2003). *Coxeter matroids*. Birkhauser.

Aichholzer, O. and Aurenhammer, F. (1996). Classifying hyperplanes in hypercubes. *SIAM J. Discrete Math.*, 9:225–232.

Bandelt, H.-J. and Pesch, E. (1989). Dismantling absolute retracts of reflexive graphs. *European J. Combinatorics*, 10:211–220.

Bell, J. L. (1986). A new approach to quantum logic. *Brit. J. Phil. Sci.*, 37:83–99.

Biacino, L. and Gerla, G. (1991). Connection structures. *Notre Dame J. Formal Logic*, 37:431–439.

Birkhoff, G. (1948). *Lattice Theory, revised edition*. Am. Math. Soc. Publications.

Birkhoff, G. and von Neumann, J. (1936). The logic of quantum mechanics. *Ann. Math.*, 37:823–843.

Björner, A., Vergnas, M. Las, Sturmfels, B., White, N., and Ziegler, G. (1993). *Oriented Matroids*. Cambridge University Press.

Bland, R. G. and Vergnas, M. Las (1978). Orientability of matroids. *J. Combin. Theory Ser. B*, 24(1):94–123.

Brandstädt, A., Le, V., and Spinrad, J. (1999). *Graph Classes*. SIAM Monographs in Discrete Mathematics and Applications.

Brightwell, G. (2000). Gibbs measures and dismantlable graphs. *J. Comb. Theory, Series B*, 78:141–166.

Brown, K. S. (1989). *Buildings*. Springer-Verlag.

Čech, E. (1966). *Topological Spaces*. John Wiley.

Clarke, B. (1981). A calculus of individuals based on "connection". *Notre Dame J. Formal Logic*, pages 204–218.

Clarke, B. (1985). Individuals and points. *Notre Dame J. Formal Logic*, 26: 61–75.

Coecke, B., Moore, D., and Wilce, A. (2000). Operational quantum logic:an overview. In Coecke, B., Moore, D., and Wilce, A., editors, *Current Research in Operational Quantum Logic: Algebras, Categories, Languages*. Kluwer.

Cohen, D. W. (1989). *An Introduction to Hilbert Space and Quantum Logic*. Springer-Verlag.

Cohn, A., Bennett, B., Gooday, J., and Gotts, N. (1997). RCC: a calculus for region based qualitative spatial reasoning. *GeoInformatica*, 1:275–316.

Coppel, W. (1998). *Foundations of Convex Geometry*. Cambridge University Press.

Duchet, P. (1988). Convex sets in graphs II, minimal path convexity. *J. Combin. Theory Series B*, 44:307–316.

Duchet, P. and Meyniel, H. (1983). Ensembles convexes dans les graphes I. *European J. Combinatorics*, pages 127–132.

Dvurečenskij, A. and Pulmannová, S. (2000). *New Trends in Quantum Structures*. Kluwer.

Engelking, R. (1989). *General Topology*. Heldermann, Berlin, revised edition edition.

Evako, A., Kopperman, R., and Mukhin, Y. V. (1996). Dimensional properties of graphs and digital spaces. *J. Math. Imaging and Vision*, 6:109–119.

Evako, A. V. (1994). Dimension on discrete spaces. *Int. J. Theoretical Physics*, 33:1553–1568.

Faure, C.-A. and Frölicher, A. (2000). *Modern projective geometry*. Kluwer.

Folkman, J. and Lawrence, J. (1978). Oriented matroids. *J. Combin. Theory Ser. B*, 25(2):199–236.

Foulis, D. J. (1999). A half century of quantum logic: what have we learned? In Aerts, D. and Pykacs, J., editors, *Quantum Structures and the Nature of Reality*, pages 1–36. Kluwer.

Foulis, D. J. and Randell, C. (1971). Lexicographic orthogonality. *J. Combinatorial Theory*, pages 157–162.

Georgatos, K. (2003). On indistinguishability and prototypes. *Logic J. of the IGPL*, 11:531–545.

Hell, P. and Nešetřil, J. (2004). *Graphs and Homomorphisma*. Oxford University Press.

Hurewicz, W. and Wallman, H. (1948). *Dimension Theory*. Princeton University Press.

Kalmbach, G. (1983). *Orthomodular Lattices*. Academic Press.

Khalimsky, E., Kopperman, R., and Meyer, P. (1990). Computer graphics and connected topologies on ordered sets of points. *Topology Appl.*, 35:1–17.

Knuth, D. E. (1991). *Axioms and Hulls*. Lecture Notes in Computer Science 606. Springer-Verlag.

Lane, S. Mac and Birkhoff, G. (1967). *Algebra*. Macmillan, 2nd edition.

Lawvere, F. (1991). Intrinsic co-Heyting boundaries and the Leibniz rule in certain toposes. *(Springer) Lect. Notes in Math.*, 1488:279–281.

Markopoulou, F. and Smolin, L. (1997). Causal evolution of spin networks. *Nucl. Phys.B508*, page 409.

Martin, N. N. and Pollard, S. (1996). *Closure Spaces and Logic*. Kluwer.

Oxley, J. (1992a). Infinite matroids. In White, N., editor, *Matroid Applications*, pages 73–90. Cambridge University Press.

Oxley, J. (1992b). *Matroid Theory*. Oxford University Press.

Pagliani, P. (1998). Intrinsic co-Heyting boundaries and informatin incompleteness in rough set analysis. In Polkowski, L. and Skowron, A., editors, *RSTC'98*, volume 1424 of *LNAI*, pages 123–130.

Penrose, R. (1971). Angular momentum: an approach to combinatorial space-time. In *Quantum Theory and Beyond*. Cambridge University Press.

Pfaltz, J. L. (1996). Closure lattices. *Discrete Mathematics*, 154:217–236.

Poincaré, H. (1905). *La Valeur de la Science*. Flammarion, Paris.

Poston, T. (1971). *Fuzzy Geometry*. PhD thesis, University of Warwick.

Pratt, I. and Lemon, O. (1997). Ontologies for plane, polygonal mereotopology. *Notre Dame J. Formal Logic*, 38(2):225–245.

Prenowitz, W. and Jantosciak, J. (1979). *Join Geometries*. Springer-Verlag.

Pták, P. and Pulmannová, S. (1991). *Orthomodular Structures as Quantum Logics*. Kluwer.

Pultr, A. (1963). An analogon of the fixed-point theorem and its application for graphs. *Comment.Math. Univ. Carol.*

Quilliot, A. (1983). *Homomorphismes, points fixes, rétractions et jeux de poursuite dans les graphes*. PhD thesis, Paris.

Randell, D., Cui, Z., and Cohn, A. (1992). A spatial logic based on regions and connection. In *Proc. 2nd Int. Conf. on Knowledge Representation and Reasoning*, pages 165–176.

Reyes, G. and Zolfaghari, H. (1996). Bi-Heyting algebras, toposes and modalities. *J. Philos. Logic*, 25.

Ronan, M. (1989). *Lectures on Buildings*. Academic Press.

Rosenfeld, A. (1986). "Continuous" functions on digital pictures. *Pattern Recog. Letters*, 4:177–184.

Roy, A. and Stell, J. (2002). A qualitative account of discrete space. In *Proc. 2nd Int. Conf. on Geographic Information Science*, volume 2478 of *LNCS*, pages 276–290. Springer.

Smolin, L. (2001). *Three Roads to Quantum Gravity*. Weidenfeld and Nicholson.

Smyth, M. B. (1995). Semi-metrics, closure spaces and digital topology. *Theoret. Comput. Sci.*, 151:257–276.

Smyth, M. B. (1997). Topology and tolerance. *Electron. Notes in Theoret. Comput. Sci.*, 6.

Smyth, M. B. (2000). Region-based discrete geometry. *J. Universal Comput. Sci.*, 6:447–459.

Smyth, M. B. and Tsaur, R. (2001–2002). Hyperconvex semi-metric spaces. *Topology Proceedings*, 26:791–810.

Smyth, M. B. and Webster, J. (2002). Finite approximation of stably compact spaces. *Applied General Topology*, 3:1–28.

Sorkin, R. (2002). Causal sets: discrete gravity. In Gomberoff, A. and Marolf, D., editors, *Valdivia Summer School, 2002 (to appear)*.

Sossinsky, A. (1986). Tolerance space theory and some applications. *Acta Applicandae Math.*, 5:137–167.

Stell, J. (2000). Boolean connection algebras: a new approach to the Region-Connection Calculus. *Artificial Intelligence*, 122:111–136.

Stell, J. and Worboys, M. (1997). The algebraic structure of sets of regions. *Lect. Notes in Comp. Sci.*, 1329:163–174.

Stolfi, J. (1991). *Oriented projective geometry*. Academic Press.

Sumner, R. (1974). Dacey graphs. *J. Australian Math. Soc.*, 18:492–502.

Tsaur, R. and Smyth, M. (2001). "Continuous" multifunctions in discrete spaces, with applications to fixed point theory. In Bertrand, G., Imiya, A., and Klette, R., editors, *Digital and Image Geometry*, volume 2243 of *LNCS*, pages 75–88. Springer.

Tsaur, R. and Smyth, M. (2004). Convexity in Helly graphs. In *MFCSIT 2004 (to appear)*.

van de Vel, M. (1993). *Theory of Convex Structures*. Elsevier, Amsterdam.

Vergnas, M. Las (1980). Convexity in oriented matroids. *J. Combin. Theory Ser. B*, 29(2):231–243.

Webster, J. (1997). *Topology and measure theory in the digital setting: on the approximation of spaces by inverse sequences of graphs*. PhD thesis, Imperial College.

Webster, R. (1995). *Convexity*. Oxford University Press.

Whitney, H. (1935). On the abstract properties of linear dependence. *American Journal of Mathematics*, 57:509–533.

Wilce, A. (2004). Topological test spaces. *To appear in: Int. J. Theor. Physics*.

Zeeman, E. C. (1962). The topology of the brain and visual perception. In Fort, M. K., editor, *Topology of 3-manifolds*. Prentice Hall, NJ.

Ziegler, G. M. (1995). *Lectures on Polytopes*. Springer.

Chapter 13

REAL ALGEBRAIC GEOMETRY AND CONSTRAINT DATABASES

Floris Geerts
Hasselt University, Transnational University of Limburg & University of Edinburgh

Bart Kuijpers
Hasselt University & Transnational University of Limburg

Second Reader

Peter Revesz
University of Nebraska–Lincoln

1. From the relational database model to the constraint database model

The constraint database model can be seen as a generalization of the classical relational database model that was introduced by Codd in the 1970s to deal with the management of alpha-numerical data, typically in business applications (Codd, 1970). A relational database can be viewed as a finite collection of tables or relations that each contain a finite number of tuples.

Fig. 13.1 shows an instance of a relational database that contains the two relations **Beer** and **Pub**. This database contains tourist information about beers and the pubs where they are served. It also contains the location of the pubs, given in (x, y)-coordinates on some tourist map. Each relation contains a finite number of tuples. A relational database is usually modeled following a database *schema*. A schema contains information on the relation names and on the names of the attributes appearing in relation. In this example, the attributes of **Beer** are *Name*, *Pub*, *City* and *Postal code*. The complete schema of the relational database of Fig. 13.1 could be written as **Beer**(*Name*, *Pub*, *City*, *Postal code*), **Pub**(*Pub*, x, y).

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 799–856.
© 2007 *Springer.*

Beer

Name	*Pub*	*City*	*Postal code*
Duvel	De Muze	Antwerpen	2000
Hoegaarden	Villicus	Hasselt	3500
Geuze	La Bécasse	Brussel	1000
...	...	...	...

Pub

Pub	x	y
De Muze	16	10
Villicus	16.1	14
La Bécasse	10.4	12.3
...	...	...

Figure 13.1. An example of a relational database consisting of the two relations **Beer** and **Pub**.

The x and y attributes of the relation **Pub** have a geometric or geographic interpretation. But values of these attributes can simply be stored as numbers, as is usually done in business databases. A tourist could consult this database to find out the locations of pubs where his/her preferred beers are served. First-order logic based languages (and their commercial versions, such as SQL) are used in the relational database model, to formulate queries like this. The vocabulary of these logics typically contains the relation names appearing in the schema of the input database. For instance, the first-order formula

$$\varphi(x, y) = \exists p \exists c \exists p' (\mathbf{Beer}(\text{Westvleteren}, p, c, p') \wedge \mathbf{Pub}(p, x, y))$$

when interpreted over the database of Fig. 13.1, defines the (x, y)-coordinates of the location of the pubs where they serve my favorite beer.

But a tourist is usually also given more explicit geographic information, e.g., in the form of maps such as the one depicted in Fig. 13.2 and he/she typically wants to ask questions that combine spatial and alpha-numeric information, such as "*Where in Flanders, not too far from the river Scheldt, can I drink a Duvel?*"

In the relational database model, it is difficult to support queries like this one. Unlike the locations of pubs, the locations of rivers or regions would require the storage of infinitely many x- and y-coordinates of points. Storing infinitely many tuples is not possible and in computer science it is customary to find *finite* representations of even infinite sets or objects.

In the 1980s, extensions of the relational model were proposed with special-purpose data types and operators. Data types like "polyline" and "polygon" were introduced to support, e.g, the storage of rivers and regions. Ad-hoc operations like intersection of polygons were added to popular query languages such as SQL. Since then, spatial database theory and technology has developed

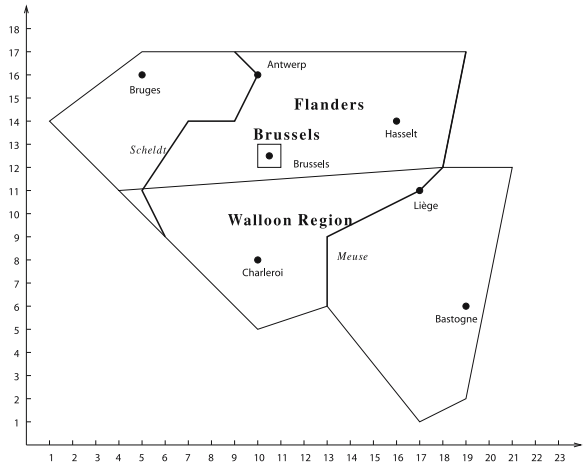


Figure 13.2. Spatial information map of Belgium.

towards more sophisticated data models and more elegant query formalisms supported by, for example, appropriate indexing techniques. For an overview of the developments in spatial databases in the last two decades, we refer to Rigaux et al., 2000.

Looking again at the polylines and polygons in Fig. 13.2, we may remark that there are other finite ways to store them, besides the indirect method of storing their corner points. Indeed, each line segment can be described by linear equations (equalities and inequalities). Moreover, polygonal figures can be described by combinations of linear inequalities. This description is more explicit than listing the corner points. If we agree that the combinations of linear equations may appear in the tuples of the relations of a database under a *geometric* attribute name, the spatial information displayed on the map of Belgium, which could be categorized into region, city, and river information, could be captured in a database with the three relations **Regions**, **Cities**, **Rivers**. Each of these relations has *Name* and *Geometry* as attributes, where the latter can be viewed as having an x-and a y-component. *Name* is a traditional alphanumeric attribute and *Geometry* has a spatial or geometric interpretation. Of course we could include more thematic information, e.g., we could add to the **City** relation the number of inhabitants.

The database instance with this schema, corresponding, to the map shown in Fig. 13.2, is given in Fig. 13.3.

Cities

Name	*Geometry*(x, y)
Antwerp	$(x = 10) \wedge (y = 16)$
Bastogne	$(x = 19) \wedge (y = 6)$
Bruges	$(x = 5) \wedge (y = 16)$
Brussels	$(x = 10.5) \wedge (y = 12.5)$
Charleroi	$(x = 10) \wedge (y = 8)$
Hasselt	$(x = 16) \wedge (y = 14)$
Liège	$(x = 17) \wedge (y = 11)$

Rivers

Name	*Geometry*(x, y)
Meuse	$\big((y \leq 17) \wedge (5x - y \leq 78) \wedge (y \geq 12)\big) \vee$
	$\big((y \leq 12) \wedge (x - y = 6) \wedge (y \geq 11)\big) \vee$
	$\big((y \leq 11) \wedge (x - 2y = -5) \wedge (y \geq 9)\big) \vee$
	$\big((y \leq 9) \wedge (x = 13) \wedge (y \geq 6)\big)$
Scheldt	$\big((y \leq 17) \wedge (x + y = 26) \wedge (y \geq 16)\big) \vee$
	$\big((y \leq 16) \wedge (2x - y = 4) \wedge (y \geq 14)\big) \vee$
	$\big((x \leq 9) \wedge (x \geq 7) \wedge (y = 14)\big) \vee$
	$\big((y \leq 14) \wedge (-3x + 2y = 7) \wedge (y \geq 11)\big) \vee$
	$\big((y \leq 11) \wedge (2x + y = 21) \wedge (y \geq 9)\big)$

Regions

Name	*Geometry*(x, y)
Brussels	$(y \leq 13) \wedge (x \leq 11) \wedge (y \geq 12) \wedge (x \geq 10)$
Flanders	$(y \leq 17) \wedge (5x - y \leq 78) \wedge (x - 14y \leq -150) \wedge$
	$(x + y \geq 45) \wedge (3x - 4y \geq -53) \wedge \big(\neg\big((y \leq 13) \wedge$
	$(x \leq 11) \wedge (y \geq 12) \wedge (x \geq 10)\big)\big)$
Walloon Region	$\big((x - 14y \geq -150) \wedge (y \leq 12) \wedge (19x + 7y \leq 375) \wedge$
	$(x - 2y \leq 15) \wedge (5x + 4y \geq 89) \wedge (x \geq 13)\big) \vee$
	$\big((-x + 3y \geq 5) \wedge (x + y \geq 45) \wedge$

Figure 13.3. Representation of the spatial database of Belgium shown in Fig. 13.2.

The geometric components of the relations in Fig. 13.3 are described using linear equalities, linear inequalities and Boolean combinations thereof, i.e., using $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation). Figures that can be described in this way are sometimes referred to as *semi-linear set* figures.

One of the most important application areas of spatial databases is Geographic Information Systems (GIS), where in most cases polygonal-shaped geometric figures are considered. In most cases this data resides in the two-dimensional plane or in the three-dimensional space (Rigaux et al., 2000). Indeed, in GIS, information is mostly linear in nature, but in other applications, like CAD-CAM, or medical imaging we can find spatial figures that are not linear. Using *polynomial* equalities and inequalities rather than just linear ones gives us wider modeling capabilities. Fig. 13.4 gives an example of a figure in the plane that can be described by the following combination of polynomial (in)equalities:

$$\begin{aligned}&(x^2/25 + y^2/16 \leq 1) \wedge (x^2 + 4x + y^2 - 2y \geq -4)\\ &\quad \wedge (x^2 - 4x + y^2 - 2y \geq -4) \wedge \big((x^2 + y^2 - 2y \neq 8) \vee (y > -1)\big).\end{aligned}$$

This figure is described by a formula containing two variables, namely x and y, representing the coordinates of points in $\mathbb{R}^2$.

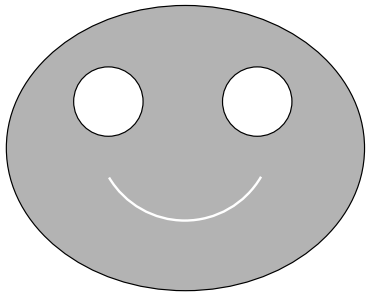


Figure 13.4. An example of a semi-algebraic set in $\mathbb{R}^2$.

Figures that can be modeled by polynomial inequalities are known, in mathematics, as *semi-algebraic sets* and their geometric and topological properties are well-studied in real algebraic geometry (Bochnak et al., 1998).

Semi-algebraic sets are, together with classical alpha-numeric data, the basic ingredients in *constraint databases*. As we have seen above in Fig. 13.2, these sets appear in a constraint database by means of a defining formula. In this sense, the constraint database model is a generalization of the relational database model.

Like the classical relational database model, first-order logic can be used to formulate queries in the constraint model. Semi-algebraic sets are described

by Boolean combinations of (linear) polynomial inequalities, which are basically quantifier-free formulas in first-order logic over the reals. This logic has addition and multiplication as functions, order as relation and zero and one as constants. In the constraint database model, an extension of this logic with predicates to address the relations in the input database is used as a basic logical query language. This logic turns out to be a language in which a lot of relevant spatial database queries can be formulated. For example, the query "*Where in Flanders, not too far from the river Scheldt, can I drink a Duvel?*" can be expressed by the formula

$$\begin{aligned}\varphi(x,y) = {} & \mathbf{Regions}(\text{Flanders}, x, y) \wedge \\ & \exists x' \exists y' (\mathbf{Rivers}(\text{Scheldt}, x', y') \wedge (x-x')^2 + (y-y')^2 < 1) \wedge \\ & \qquad \exists p \exists c \exists p' (\mathbf{Pubs}(p, x, y) \wedge \mathbf{Beer}(\text{Duvel}, p, c, p')).\end{aligned}$$

Here, we translate "not to far from the river Scheldt" by "at most distance 1 from the some point of the Scheldt". We remark that some variables in this expression are assumed to range over finite domains (namely p, c, p'), but others range over the real numbers (namely x, y, x' and y'). Nevertheless, it turns out that queries expressed by first-order formulas like this one can be effectively evaluated on constraint databases. In our example the output is a two-dimensional geometric object and the query evaluation algorithm guarantees that it can also be described by a Boolean combination of polynomial inequalities.

The ideas presented above are at the basis of the constraint database model. The basic idea is to extend or generalize the relational model and not only to allow finite relations, but also finitely representable relations.

We remark that the constraint database model was introduced by Kanellakis et al., 1995. It has received a lot of research attention since. An overview of research results in this field can be found in Kuper et al., 2000, and Revesz has written a textbook on the subject (Revesz, 2002).

Overview. This chapter is organized as follows. In Sec. 2, we describe the constraint database model with its data models and basic query languages. Sec. 3 gives an overview of some definitions and results in real algebraic geometry that will be used further on. In Sec. 4, we discuss query evaluation in the constraint database model through quantifier elimination. We also outline some quantifier elimination algorithms there. Sec. 5 is devoted to the expressive power of first-order logic over the reals as a query language for constraint databases. Topological queries get special attention. Finally, in Sec. 6, we discuss some more powerful query languages for constraint databases that are extensions of first-order logic, with transitive closure operators, with while-loop and with topological operators.

2. Constraint data models and query languages

In this section, we define the logics $\mathrm{FO}(+, \times, <, 0, 1)$, i.e., *first-order logic with polynomial constraints*, and $\mathrm{FO}(+, <, 0, 1)$, i.e., *first-order logic with linear constraints*, and show how they form the basis of the constraint approach in both the modeling and querying of spatial data. More specifically, we introduce the *polynomial* and *linear constraint model* and extend $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$ to query languages for the respective models.

2.1 The logics $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$

Let $(+, \times, <, 0, 1)$ be a so-called *vocabulary* with two functions symbols of arity two ($+$ and $\times$), one predicate symbol of arity two ($<$), and two constant symbols (0 and 1). In the constraint model, this vocabulary will be interpreted on the *real field*, i.e., the structure consisting of the set of real numbers, $\mathbb{R}$, equipped with the standard addition, multiplication, and order.

We define $\mathrm{FO}(+, \times, <, 0, 1)$ as the *first-order logic over the vocabulary* $(+, \times, <, 0, 1)$. We build formulas in $\mathrm{FO}(+, \times, <, 0, 1)$ in the standard way: a *term* t in $\mathrm{FO}(+, \times, <, 0, 1)$ is either a *variable* x_i; a constant (0 or 1); or of the form $t + t'$ or $t \times t'$ for terms t and t'. In other words, terms are polynomials with integer coefficients. Next, *atomic formulas* in $\mathrm{FO}(+, \times, <, 0, 1)$ are formulas of the form $t = t'$ or $t < t'$ for terms t and t'. Finally, formulas in $\mathrm{FO}(+, \times, <, 0, 1)$ are built from atomic formulas by using the *Boolean connectives* ($\wedge$, $\vee$, or $\neg$) and quantifiers ($\forall x_i$ or $\exists x_i$). A variable is called *free* in a formula if it is not bounded by a quantifier. We denote by $\varphi(x_1, \ldots, x_n)$ the fact that the $\mathrm{FO}(+, \times, <, 0, 1)$ formula φ has n free variables $x_1, \ldots, x_n$. A formula without any free variables is called a *sentence*. A formula without quantifiers is called *quantifier-free*.

Similarly, we define $\mathrm{FO}(+, <, 0, 1)$ as the restriction of $\mathrm{FO}(+, \times, <, 0, 1)$ in which formulas are constructed from terms which do not use multiplication (i.e., formulas without $\times$). In other words, the terms in $\mathrm{FO}(+, <, 0, 1)$ are polynomials with integer coefficients of degree at most one. We also say that $\mathrm{FO}(+, <, 0, 1)$ *is the first-order logic over the vocabulary* $(+, <, 0, 1)$.

We define the *satisfaction* of a formula $\varphi(x_1, \ldots, x_n)$ in $\mathrm{FO}(+, \times, <, 0, 1)$ by real numbers $r_1, \ldots, r_n \in \mathbb{R}$, denoted by

$$(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, \ldots, r_n),$$

inductively on the structure of φ:

- $(\mathbb{R}, +, \times, <, 0, 1) \models (t = t')(r_1, \ldots, r_n)$ if $t(r_1, \ldots, r_n) = t'(r_1, \ldots, r_n)$;
- $(\mathbb{R}, +, \times, <, 0, 1) \models (t < t')(r_1, \ldots, r_n)$ if $t(r_1, \ldots, r_n) < t'(r_1, \ldots, r_n)$;

- $(\mathbb{R}, +, \times, <, 0, 1) \models (\neg\varphi)(r_1, \ldots, r_n)$ if $(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, \ldots, r_n)$ does not hold;
- $(\mathbb{R}, +, \times, <, 0, 1) \models (\varphi \wedge \psi)(r_1, \ldots, r_n)$ if $(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, \ldots, r_n)$ and $(\mathbb{R}, +, \times, <, 0, 1) \models \psi(r_1, \ldots, r_n)$;
- $(\mathbb{R}, +, \times, <, 0, 1) \models (\varphi \vee \psi)(r_1, \ldots, r_n)$ if $(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, \ldots, r_n)$ or $(\mathbb{R}, +, \times, <, 0, 1) \models \psi(r_1, \ldots, r_n)$;
- $(\mathbb{R}, +, \times, <, 0, 1) \models (\forall x_n \varphi)(r_1, \ldots, r_{n-1})$ if for all elements $r \in \mathbb{R}$, $(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, \ldots, r_{n-1}, r)$; and
- $(\mathbb{R}, +, \times, <, 0, 1) \models (\exists x_n \varphi)(r_1, \ldots, r_{n-1})$ if there exists an element $r \in \mathbb{R}$, $(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, \ldots, r_{n-1}, r)$.

As described above, in the constraint model, the satisfaction of formulas in $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$ is defined with respect to the real field $\mathbb{R}$. However, any mathematical structure which interprets the vocabularies $(+, \times, <, 0, 1)$ or $(+, <, 0, 1)$ can be used instead.

Of particular importance in the constraint model are the quantifier-free formulas in $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$. As we will see in the next section, the representation of spatial objects by means of quantifier-free formulas is the basis of the data model in constraint databases. In Sec. 4, we show that both $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$ admit *quantifier elimination*. In short, this means that any formula in $\mathrm{FO}(+, \times, <, 0, 1)$ (respectively $\mathrm{FO}(+, <, 0, 1)$) is equivalent to a quantifier-free formula in $\mathrm{FO}(+, \times, <, 0, 1)$ over $\mathbb{R}$ (respectively in $\mathrm{FO}(+, <, 0, 1)$). Hence, we do not loose any generality by considering quantifier-free formulas only. As mentioned in the introduction, a (quantifier-free) formula represents a possibly infinite set of points. More specifically, they describe sets of points which correspond to *semi-algebraic sets*, in case of $\mathrm{FO}(+, \times, <, 0, 1)$, and *semi-linear sets*, in case of $\mathrm{FO}(+, <, 0, 1)$(see also Sec. 3).

EXAMPLE 13.1 In Fig. 13.4 of Sec. 1, the smiling face shows all pairs $(r_1, r_2) \in \mathbb{R}^2$ that satisfy $\varphi(x, y)$, i.e., $(\mathbb{R}, +, \times, <, 0, 1) \models \varphi(r_1, r_2)$, where $\varphi(x, y)$ is the quantifier-free formula

$$x^2/25 + y^2/16 \leq 1 \wedge x^2 + 4x + y^2 - 2y \geq -4 \wedge \\ x^2 - 4x + y^2 - 2y \geq -4 \wedge (x^2 + y^2 - 2y \neq 8 \vee y > -1).$$

We remark that φ has two free variables and that it uses polynomials of degree at most two.

Apart from the modeling of spatial data, the logics $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$ serve also as the basis of the standard *query languages* in the constraint model. We come back to this point in the next sections.

2.2 The polynomial constraint data model

First, we discuss the general polynomial constraint model which is based on $\mathrm{FO}(+, \times, <, 0, 1)$. In the next section, we elaborate on the linear constraint model which uses $\mathrm{FO}(+, <, 0, 1)$.

The polynomial constraint data model. A *database schema* $\mathcal{S}$ is a finite set $\{S_1, \ldots, S_k\}$ of relation names. Each relation name S_i ($i = 1, \ldots, k$) is of some arity n_i, which is an integer. A *polynomial constraint relation instance* of S_i, or *constraint relation* of S_i for short, maps S_i to a quantifier-free formula $\varphi_{S_i}(x_1, \ldots, x_{n_i})$ with n_i free variables in the logic $\mathrm{FO}(+, \times, <, 0, 1)$. A *(polynomial) constraint database instance over* $\mathcal{S}$ consists of a set of constraint relations of $S_1, \ldots, S_k$.

The *semantics* of a relation instance of S_i, denoted by $I(S_i)$, is the possibly infinite (semi-algebraic) subset

$$\{(r_1, \ldots, r_{n_i}) \in \mathbb{R}^{n_i} \mid (\mathbb{R}, +, \times, <, 0, 1) \models \varphi_{S_i}(r_1, \ldots, r_{n_i})\}.$$

The semantics of a (polynomial) constraint database instance D, denoted by $I(D)$, over the schema $\mathcal{S}$, is the collection of semi-algebraic sets $I(S_i)$, with S_i a relation name appearing in $\mathcal{S}$.

EXAMPLE 13.2 Let $\mathcal{S} = \{S\}$, where S is a binary relation name. A constraint database instance D over $\mathcal{S}$ maps, for instance, S to the quantifier-free $\varphi(x, y)$ given in Example 13.1. The semantics of D is the smiling face shown in Fig. 13.4.

It is clear that the same semi-algebraic set can be represented by different formulas. Indeed, consider again Fig. 13.4. Suppose that the description of the smiling face given in Example 13.1 is extended with the disjunct $(x = 0 \wedge -1/2 \leq y \leq 1/2)$ (i.e., a vertical line segment), representing a nose. This new representation will not lead to the addition of new points in the smiling face, since all the points in the nose are already part of the face.

Two constraint relations of S and S' (i.e., formulas) are said to be *equivalent* if $I(S)$ and $I(S')$ are the same semi-algebraic set (i.e., if $I(S) = I(S')$). Similarly, we say that two database instances are equivalent if their relations are pairwise equivalent.

REMARK 13.3 In the remainder of this chapter, we use the terms *constraint relation* and *semi-algebraic set* interchangeably, since these notions refer to the same objects, albeit from different perspectives.

Database queries in the constraint model. Before we explain how to use $\mathrm{FO}(+, \times, <, 0, 1)$ as a query language for the polynomial constraint model, we define what we mean by a query on a constraint database. In standard relational databases, a query is a (partial) function associating with each input database

instance an output relation instance. In the constraint setting, however, there are two ways of looking at a query.

- First, as in the relational setting, we can define a *k-ary query over a database schema $\mathcal{S}$* as a partial function which associates with a database instance $I(D)$ (i.e., a collection of semi-algebraic sets), a semi-algebraic set in $\mathbb{R}^k$, where D is any database instance of $\mathcal{S}$.
- Second, we can also view a k-ary query over $\mathcal{S}$ as a partial function associating with each database instance D (i.e., a collection of quantifier-free formulas), a quantifier-free formula in $\mathrm{FO}(+, \times, <, 0, 1)$ with k free variables.

We call the first type of query an *unrestricted query*; the second is called a *constraint query*. A constraint query clearly only makes sense if it maps two equivalent database instances to equivalent relation instances (i.e., equivalent quantifier-free $\mathrm{FO}(+, \times, <, 0, 1)$-formulas). If a constraint query satisfies this property, we call a constraint query *consistent*. We remark that a consistent constraint query corresponds to a *unique* unrestricted query.

EXAMPLE 13.4 Let $\mathcal{S}$ consist of binary relation S. Consider the constraint query Q which maps any constraint relation of S, given by φ_S, to the highest degree of polynomials appearing in φ_S. This query is clearly not consistent. Indeed, let $\varphi_S \equiv x^2 + y^2 = 1$ and $\varphi'_S \equiv (x^2 + y^2)^2 = 1$. Both formulas correspond to the same semi-algebraic set, i.e., the standard circle of radius 1. In contrast, Q returns 2 on input φ_S, whereas it returns 4 on input φ'_S.

In the following, when we refer to a *constraint database query*, we mean a consistent constraint query.

The logic $\mathrm{FO}(+, \times, <, 0, 1)$ as a query language for polynomial constraint databases. In this section, we take a closer look at the standard query language for polynomial constraint databases, which is an extension of $\mathrm{FO}(+, <, 0, 1)$ with predicates to address constraint relations that appear in the input database.

If we consider queries over a database input schema $\mathcal{S} = \{S_1, \ldots, S_k\}$, then we can associate a query with a formula in the first-order logic over the vocabulary $(+, \times, <, 0, 1, S_1, \ldots, S_k)$. Let $\varphi(x_1, \ldots, x_m)$ be such a formula over$(+, \times, <, 0, 1, S_1, \ldots, S_k)$. Given a constraint database D over $\mathcal{S}$, we interpret $\varphi(x_1, \ldots, x_m)$ over the $(\mathbb{R}, +, \times, <, 0, 1)$, extended with the semi-algebraic sets, $I(S_1), \ldots, I(S_k)$ as given by D. More specifically, the m-ary answer set of $\varphi(x_1, \ldots, x_m)$ is defined as

$$\{(r_1, \ldots, r_m) \in \mathbb{R}^m \mid (\mathbb{R}, +, \times, <, 0, 1, I(S_1), \ldots, I(S_k)) \models \varphi(r_1, \ldots, r_m)\}.$$

We also write the above for short as

$$\{(r_1, \ldots, r_m) \in \mathbb{R}^m \mid (\mathbb{R}, D) \models \varphi(r_1, \ldots, r_m)\}.$$

It is clear that equivalent databases result in the same answer set. We say that φ *expresses* the corresponding (unique) unrestricted query. In the sequel, we refer to these extensions of $\mathrm{FO}(+, \times, <, 0, 1)$ simply by $\mathrm{FO}(+, \times, <, 0, 1)$ if the input schema is clear from the context or irrelevant.

An important property of any query language is that it is *closed*, i.e., the result of query should admit a representation in the same data model as the source relations. In particular, for $\mathrm{FO}(+, \times, <, 0, 1)$ to be closed it should be the case that the result is a quantifier-free formula in $\mathrm{FO}(+, \times, <, 0, 1)$ again. However, since $\mathrm{FO}(+, \times, <, 0, 1)$ admits quantifier-elimination, and given the way $\mathrm{FO}(+, \times, <, 0, 1)$-formulas are evaluated, this requirement is satisfied (see also Sec. 4).

In Sec. 1, we gave examples of queries expressed in $\mathrm{FO}(+, \times, <, 0, 1)$. We give some more examples here.

EXAMPLE 13.5 Let $Q_{bounded}$ be the unrestricted query which returns true if and only if the input semi-algebraic set in $\mathbb{R}^2$ is bounded. In the first-order logic over $(+, \times, <, 0, 1, S)$, where S is a binary relation name, the sentence

$$\exists \varepsilon(\varepsilon \neq 0 \wedge \forall x \forall y(S(x, y) \rightarrow x^2 + y^2 < \varepsilon^2))$$

expresses $Q_{bounded}$.

EXAMPLE 13.6 Let $Q_{interior}$ be the query that returns all points of any input semi-algebraic set in $\mathbb{R}^2$ that have a neighborhood that is completely in the semi-algebraic set. Hence, $Q_{interior}$ returns the *topological interior* of a semi-algebraic set in $\mathbb{R}^2$. This query can be expressed as

$$\exists r \forall x' \forall y'(r \neq 0) \ \wedge \ ((x - x')^2 + (y - y')^2 < r^2 \rightarrow S(x', y')).$$

We remark that this formula has two free variables, so it defines a semi-algebraic set in $\mathbb{R}^2$.

In the next section, we discuss the expressive power of the query language $\mathrm{FO}(+, \times, <, 0, 1)$. For the moment, let us merely say that it is "rather limited." Topological queries such as the topological interior are expressible in this logic, but we will see in Sec. 5 that important queries are not expressible in $\mathrm{FO}(+, \times, <, 0, 1)$. More specifically, we will see that the query that expresses that a spatial database is *topologically connected* is not expressible. Due to the importance of this query in the spatial database practice, many efforts to extend $\mathrm{FO}(+, \times, <, 0, 1)$ to richer query languages exist, some of which we discuss in Sec. 6.

2.3 The linear constraint model: an application in Geographic Information Systems

Next, we discuss the linear constraint model, which is less expressive than the polynomial constraint model (as we will illustrate in Sec. 5), but nevertheless

powerful enough to model applications like Geographic Information Systems, or GIS for short.

The linear constraint data model. Since we emphasize the GIS aspect of the linear model here, we will also combine linear spatial information with classical alpha-numeric information, as is customary in the GIS practice. Therefore, for the sake of illustrating the suitability for GIS, we consider more general database schemas and instances in this section.

Also, the linear constraint database model can be seen as based on the relational model. Moreover, linear constraint databases also require a lot of traditional database capabilities. In particular, if the linear constraint database consists purely of non-spatial flat relations, it degenerates into a traditional database for which the relational model offers a well-accepted representation.

More formally, a *linear constraint database scheme* $\mathcal{S}$ consists of a finite set of relation names $S_1, \ldots, S_k$. Each relation name S_i ($i = 1, \ldots, k$) is of some type $[n_i, m_i]$, with n_i and m_i integers. A *linear constraint database instance* is a mapping that assigns a linear relation instance to each relation name appearing in the database scheme. A *linear relation instance* of S_i, also called a *linear relation* for short, is a finite set of linear tuples of type $[n_i, m_i]$. A *linear tuple* of type $[n_i, m_i]$ is straightforwardly defined as a tuple of the form

$$(c_1, \ldots, c_{n_i}, \varphi(x_1, \ldots, x_{m_i}))$$

where $c_1, \ldots, c_{n_i}$ are thematic values, typically from some alpha-numeric domain U (for instance, U could be the set of all strings over our alphabet and natural numbers) and $\varphi(x_1, \ldots, x_{m_i})$ is a quantifier-free formula in the logic $\mathrm{FO}(+, <, 0, 1)$ with m_i free variables.

The semantics of a linear tuple $t = (c_1, \ldots, c_{n_i}, \varphi(x_1, \ldots, x_{m_i}))$ of type $[n_i, m_i]$ is the possibly infinite subset of $U^{n_i} \times \mathbb{R}^{m_i}$ defined as the Cartesian product $\{(c_1, \ldots, c_{n_i})\} \times A_i$, in which $A_i \subseteq \mathbb{R}^{m_i}$ is the semi-linear set

$$\{(r_1, \ldots, r_{m_i}) \in \mathbb{R}^{m_i} \mid (\mathbb{R}, +, <, 0, 1) \models \varphi(r_1, \ldots, r_{m_i})\}.$$

This subset of $U^{n_i} \times \mathbb{R}^{m_i}$ can be interpreted as a possibly infinite $(n_i + m_i)$-ary relation, denoted $I(t)$. The semantics of a linear relation, S_i, denoted $I(S_i)$, is defined as $I(S_i) = \bigcup_{t \in S_i} I(t)$. Finally, the semantics of a linear spatial database, D over the schema $\mathcal{S}$, is the set of relations $I(S_i)$ with S_i a linear relation name appearing in the schema $\mathcal{S} = \{S_1, \ldots, S_k\}$ of D.

For GIS, where spatial information is often modeled in either the *vector model* or the *raster model*, and combined with traditional alpha-numeric information often stored in a relational database, the linear constraint model is powerful enough. Indeed, in the vector model, three types of planar spatial objects are standardly used, namely *points*, *polylines* and *polygons*. In the raster

model, the plane $\mathbb{R}^2$ is divided by a regular grid. Clearly, if we assume the grid to be finite, both types of data can be modeled in the linear constraint model.

EXAMPLE 13.7 In Sec. 1, we introduced the example of a map containing infomation about Belgium, as illustrated in Fig. 13.2. The spatial information displayed on the map of Belgium can be categorized into city, river, and region information. Therefore, we introduced three relations, each containing one of these spatial information sources (Fig. 13.3). The relations **Cities**, **Rivers** and **Regions** are of type $[1, 2]$ and model respectively points, polylines and polygons. Their thematic component contains names (or string information), whereas their spatial component contains formulas describing spatial features of Belgium. This example illustrated that the linear constraint model is suitable for GIS.

The logic $\mathrm{FO}(+, <, 0, 1)$ **as a query language for Geographic Information Systems.** In this section, we take a closer look at the standard query language for linear constraint databases which is an extension of $\mathrm{FO}(+, <, 0, 1)$ with predicates to address linear constraint relations that appear in the input database. Because of the mixed presence of thematic and spatial information, this query language will be an extension of $\mathrm{FO}(+, <, 0, 1)$ in the sense of a two-sorted logic. More specifically, if we consider queries over a database input schema $\mathcal{S} = \{S_1, \ldots, S_k\}$, we have, apart from the terms, formulas and quantifications possible in $\mathrm{FO}(+, <, 0, 1)$, the following ingredients:

- apart from (real) variables $x_1, x_2, \ldots$ ranging over $\mathbb{R}$, we also have infinitely many *thematic variables* $v_1, v_2, \ldots$ ranging over U and distinct from the set of real variables;
- we have atomic formulas of the form $v_1 = v_2$, with v_1 and v_2 thematic variables;
- we have atomic formulas of the form $S_i(v_{i_1}, \ldots, v_{i_{n_i}}; t_{j_1}, \ldots, t_{j_{m_i}})$, with S_i a relation name of type $[n_i, m_i]$, $v_{i_1}, \ldots, v_{i_{n_i}}$ are thematic variables, and $t_{j_1}, \ldots, t_{j_{m_i}}$ are terms in $\mathrm{FO}(+, <, 0, 1)$; and
- universal and existential quantification of thematic variables.

In the following, we will refer to this extension of $\mathrm{FO}(+, <, 0, 1)$, simply as $\mathrm{FO}(+, <, 0, 1)$. Similar to the case of $\mathrm{FO}(+, \times, <, 0, 1)$, a formula $\varphi(v_1, \ldots, v_n, x_1, \ldots, x_m)$ in $\mathrm{FO}(+, <, 0, 1)$ expresses a constraint query of type $[n, m]$.

Finally, we shall give some typical example queries, illustrating the expressive power of $\mathrm{FO}(+, <, 0, 1)$.

EXAMPLE 13.8 An example of a (very simple) linear spatial query on the database in Example 13.3 is *"Find all cities that lie on a river and give their*

names and the names of the rivers they lie on." This query can be expressed by the following first-order formula:

$$\varphi(c, r) = \exists x \exists y(\textbf{Cities}(c, x, y) \wedge \textbf{Rivers}(r, x, y)).$$

This formula defines an output relation of type $[2, 0]$.

In all the remaining queries, we shall assume the input database consists of one relation S of type $[0, 2]$.

EXAMPLE 13.9 The following $\mathrm{FO}(+, <, 0, 1)$-sentence expresses $Q_{bounded}$ (see Example 13.5):

$$\exists d \forall x \forall y(S(x, y) \rightarrow -d < x \wedge x < d \wedge -d < y \wedge y < d).$$

EXAMPLE 13.10 Several topological properties of a semi-linear set can be expressed in $\mathrm{FO}(+, <, 0, 1)$. For instance, the query $Q_{interior}$ (see Example 13.6) is expressed by the $\mathrm{FO}(+, <, 0, 1)$ formula

$$\varphi(x, y) = \exists \varepsilon \forall x' \forall y'(\varepsilon \neq 0) \wedge ((\mid x - x' \mid < \varepsilon \wedge \mid y - y' \mid < \varepsilon) \rightarrow S(x', y')).$$

The formula $\varphi(x, y)$ represents a semi-linear set in $\mathbb{R}^2$.

In spite of all this, $\mathrm{FO}(+, <, 0, 1)$ cannot be considered as a fully adequate query language for practical purposes. More specifically, there are very simple queries which are not expressible in $\mathrm{FO}(+, <, 0, 1)$, which are expressible in $\mathrm{FO}(+, \times, <, 0, 1)$. We return to this issue in Sec. 5.

3. Introduction to real algebraic geometry

In this section, we define and discuss semi-algebraic and semi-linear sets and review some well-known properties of these sets. We are interested in sets which are situated in the n-dimensional Euclidean space $\mathbb{R}^n$.

An excellent introduction to real-algebraic geometry can be found in (Coste, 2000b). Proofs of all the theorems given in this section can be found there. More advanced is the standard book in the field (Bochnak et al., 1987) and for a more algorithmic point of view we refer to (Basu et al., 2003a). An interesting book covering many other aspects of real algebraic geometry is (Benedetti and Risler, 1990). On a very advanced level, investigations of real-algebraic geometry in terms of constructible sets, real spectra, and spaces of orderings can be found in (Andradas et al., 1996).

Finally, the generalization of real-algebraic geometry to so-called o-minimal geometry is described in (Coste, 2000a). An excellent book on o-minimal structures is (van den Dries, 1998). Interestingly, many results from constraint databases described in this chapter can be generalized to the o-minimal setting. We refer to the standard book on constraint databases for more details (Kuper et al., 2000).

3.1 Semi-algebraic sets and their basic properties

Definition of semi-algebraic sets. A *semi-algebraic subset of* $\mathbb{R}^n$ is a subset of points $\vec{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$ satisfying a Boolean combination (expressed by disjunction, conjunction and negation—or in set-theoretic terms by union, intersection, and complement) of polynomial equations and inequalities with integer coefficients. It is easy to see that every semi-algebraic set in $\mathbb{R}^n$ is the finite union of sets of the form

$$\{\vec{x} \in \mathbb{R}^n \mid f(\vec{x}) = 0, \quad g_1(\vec{x}) > 0, g_2(\vec{x}), \ldots, g_\ell(\vec{x}) > 0\},$$

where $f, g_1, \ldots, g_\ell$ are multivariate polynomials in the variables $x_1, \ldots, x_n$ with integer coefficients. A *semi-linear subset of* $\mathbb{R}^n$ is a semi-algebraic subset which is described by multivariate polynomials of degree at most one (i.e., linear multivariate polynomials).

REMARK 13.11 It is easy to see that the class of semi-algebraic sets defined above coincides with the class of sets represented by quantifier-free formulas in $\mathrm{FO}(+, \times, <, 0, 1)$. We therefore are free to choose either of the two representations. We more often use the representation in terms of quantifier-free formulas.

EXAMPLE 13.12 In the introductory section we have already given an example of semi-algebraic sets (see, e.g., Fig. 13.4). Semi-algebraic sets can be used to model various spatial situations, but also spatio-temporal phenomena, as is illustrated in Fig. 13.5. Here a potential scene from Star Trek is depicted in which the starship Enterprise fires a photon torpedo. This scene plays in the three-dimensional (x, y, t) space, where x and y are spatial coordinates and t represents a time coordinate. The star ship remains at a constant position in space and can therefore be described by some fixed formula

$$\varphi_{\mathrm{Enterprise}}(x, y, t) = \big((x^2 + y^2 = 1) \vee (x^2 + y^2 = (1/4)^2) \vee \cdots\big)$$

in which t does not appear. A fired photon torpedo follows the dotted line (between the moments $t = 0$ and $t = 1$) an then explodes (depicted as increasing dotted circles, between $t = 1$ and $t = 2$). At the bottom of Fig. 13.5 three frames of the movie are shown: at $t = 1/2, 1$ and 2. The complete movie can be described by the set

$$\begin{aligned}\{(x, y, t) \in \mathbb{R}^2 \times \mathbb{R} \mid (\varphi_{\mathrm{Enterprise}}(x, y) \wedge (0 \leq t \leq 2)) \vee \\ ((y = 0 \wedge x = 4t) \wedge (0 \leq t \leq 1)) \vee \\ \big(((x - 4)^2 + y^2 \leq (t - 1)) \wedge (1 < t \leq 2)\big)\}.\end{aligned}$$

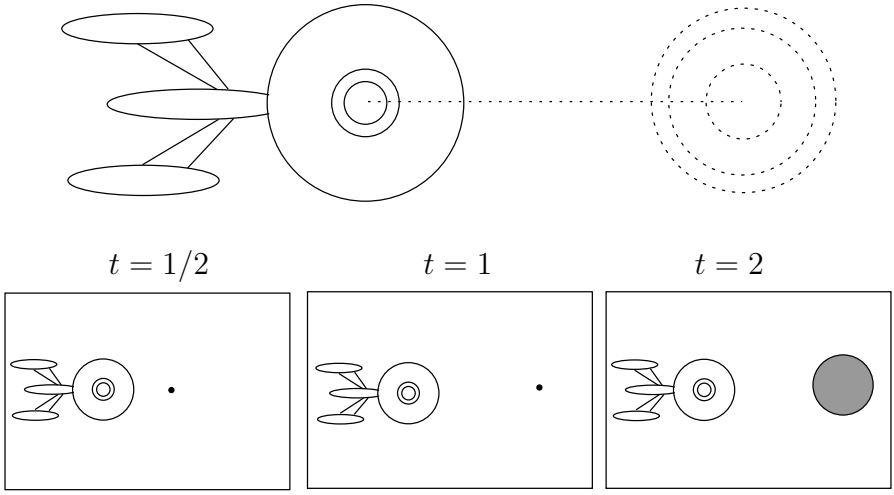


Figure 13.5. USS Enterprise firing a photon torpedo at a (cloaked) Klingon vessel.

Basic properties of semi-algebraic sets. The class of semi-algebraic sets is closed under finite unions, intersections and complements. Moreover, if $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$ are semi-algebraic, then the cartesian product $A \times B$ is a semi-algebraic subset of $\mathbb{R}^{m+n}$. A much deeper result is that the class of semi-algebraic sets is closed under projection as well:

THEOREM 13.13 (TARSKI-SEIDENBERG) *Let A be a semi-algebraic subset of $\mathbb{R}^{n+1}$ and let $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ be the projection on the first n coordinates. Then $\pi(A)$ is a semi-algebraic set of $\mathbb{R}^n$.*

One may wonder whether all sets of $\mathbb{R}^n$ are semi-algebraic. Already for $n = 1$, it can be shown that there are subsets that are not semi-algebraic. In fact, every semi-algebraic subset of $\mathbb{R}$ is known to be a finite union of open intervals (possibly unbounded) and points. Fig. 13.6 gives an example of a one-dimensional semi-algebraic set. It is the union of four open intervals (the leftmost being unbounded) and five points (three of which are isolated).

Figure 13.6. An example of a semi-algebraic set in $\mathbb{R}$.

From this property it follows that the set of natural numbers $\mathbb{N}$ is not a semi-algebraic subset of $\mathbb{R}$. Similarly, it is easily verified that the zig-zag line in $\mathbb{R}^2$ shown in Fig. 13.7 is not semi-algebraic.

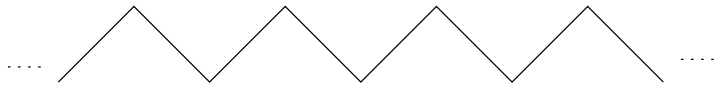


Figure 13.7. An example of a subset of $\mathbb{R}^2$ that is not semi-algebraic.

Let A be a semi-algebraic set of $\mathbb{R}^n$ and let $f : A \to \mathbb{R}$ be a real-valued function. Then f is called a *semi-algebraic function* if its graph

$$\Gamma(f) = \{(\vec{x}, r) \in A \times \mathbb{R} \mid \vec{x} \in A \text{ and } r = f(\vec{x})\}$$

is a semi-algebraic set of $\mathbb{R}^{n+1}$.

Curve selection. The following result says that any point on the border of a semi-algebraic set can be connected to the set via a continuous curve.

THEOREM 13.14 (CURVE SELECTION) *Let A be a semi-algebraic set of $\mathbb{R}^n$, and let $\vec{x} \in \overline{A} \setminus A$. Then there exists a continuous semi-algebraic function $\gamma : [0,1] \to \mathbb{R}^n$ such that $\gamma(0) = \vec{x}$ and $\gamma((0,1]) \subseteq A$.*

A proof of this theorem can be found, e.g., in (Bochnak et al., 1987, Proposition 2.5.3).

REMARK 13.15 A set A of $\mathbb{R}^n$ is *connected* if there exists no open sets U, V of $\mathbb{R}^n$ such that $A = U \cup V$, $U \cap \overline{V} = \emptyset$ and $\overline{U} \cap V = \emptyset$. A set A is semi-algebraically arc-connected if between any two points $\vec{s}, \vec{t} \in A$ there exists a semi-algebraic funtion $\gamma : [0,1] \to \mathbb{R}^n$ such that $\gamma(0) = \vec{s}$, $\gamma(1) = \vec{t}$ and $\gamma([0,1]) \subseteq A$. It can be easily verified that from the curve selection theorem, it follows that for a semi-algebraic A of $\mathbb{R}^n$ being connected coincides with being semi-algebraically arc-connected.

We remark that the curve selection theorem also holds when semi-algebraic is replaced by semi-linear.

3.2 Decompositions of semi-algebraic sets

Topological decomposition. The semi-algebraic sets of $\mathbb{R}^1$ are characterized above as being finite unions of open intervals and points. A similar characterization exists for semi-algebraic sets of $\mathbb{R}^n$, which we state here. A proof of this result can be found in (Bochnak et al., 1987, Theorem 2.3.6). We first recall the definition of a homeomorphism: a *homeomorphism* h between two sets X and Y is a continuous bijection which has continuous inverse. Two sets are called *homeomorphic* if there exists a homeomorphism between them.

THEOREM 13.16 *Let A be a semi-algebraic subset of $\mathbb{R}^n$. Then A can be written as a finite union*

$$A = \bigcup_{i=0}^{n} \bigcup_{j=1}^{m_i} A_{ij},$$

where each A_{ij} is homeomorphic to the open cube $(0,1)^i$.

We remark that the *dimension of* $(0,1)^i$ is i. So, this theorem states that any semi-algebraic set of $\mathbb{R}^n$ can be decomposed into finite unions of objects that are from a topological point of view, open cubes of dimension lower or equal to n.

Cylindrical algebraic decomposition. In practice, more refined decompositions of semi-algebraic sets are used that are also computable by more or less efficient algorithms. One such decomposition is given by the *cell decomposition theorem* for semi-algebraic sets. Before we can state this theorem, we will need the notion of cylindrical algebraic decomposition (CAD) of $\mathbb{R}^n$: A CAD of $\mathbb{R}^n$ is a special partition of $\mathbb{R}^n$ into finitely many cells. The definition is by induction on n:

(i) a CAD of $\mathbb{R}^1$ is a collection

$$\big\{(-\infty, a_1), (a_1, a_2), \ldots, (a_k, +\infty), \{a_1\}, \ldots, \{a_k\}\big\},$$

of open intervals and points, where $a_1 < \cdots < a_k$ are points in $\mathbb{R}$.

(ii) a CAD of $\mathbb{R}^{n+1}$ is a finite partition of $\mathbb{R}^{n+1}$ into (semi-algebraic) *cells* A such that the set of projections $\pi(A)$ is again a CAD of $\mathbb{R}^n$. Here, $\pi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is again the usual projection map defined by $\pi(x_1, \ldots, x_n, x_{n+1}) = (x_1, \ldots, x_n)$.

We still have to specify what a cell in $\mathbb{R}^{n+1}$ is. Let $(i_1, \ldots, i_m)$ be a sequence of zeros and ones of length m. We define a cell inductively on m as follows:

(i) a (0)-cell is a one-element set $\{r\}$ of $\mathbb{R}$, a (1)-cell is an open interval $(a, b) \subseteq \mathbb{R}$.

(ii) Suppose $(i_1, \ldots, i_m)$-cells are already defined. Then an $(i_1, \ldots, i_m, 0)$-cell is the graph $\Gamma(f)$ of a continuous semi-algebraic function $f{:}X{\to}\mathbb{R}$, where X is an $(i_1, \ldots, i_m)$-cell. Furthermore, an $(i_1, \ldots, i_m, 1)$-cell is a set of the form

$$(f, g)_X = \big\{(\vec{x}, r) \in X \times \mathbb{R} \mid \vec{x} \in X \text{ and } f(\vec{x}) < r < g(\vec{x})\big\},$$

where X is an $(i_1, \ldots, i_m)$-cell and f, g are continuous semi-algebraic functions on X, possibly equal to the constant functions $+\infty$ or $-\infty$.

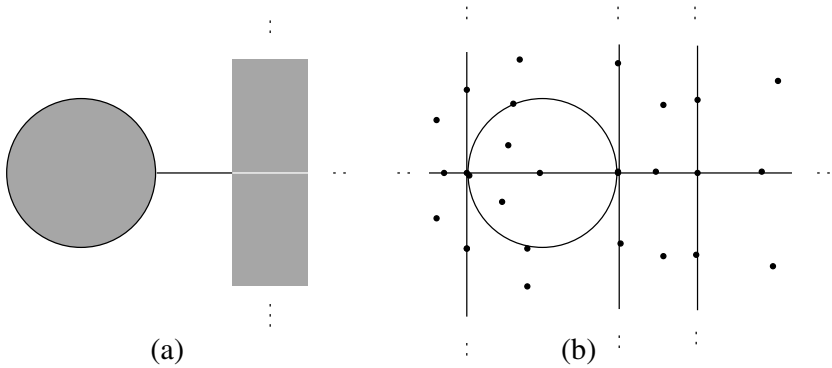


Figure 13.8. An example of a CAD in $\mathbb{R}^2$.

A *cell in* $\mathbb{R}^n$ is an $(i_1, \ldots, i_n)$-cell for some sequence $(i_1, \ldots, i_n)$. A semi-algebraic A of $\mathbb{R}^n$ is said to be *partitioned* by a CAD $\mathcal{D}$ of $\mathbb{R}^n$ if each cell in $\mathcal{D}$ is either part of or disjoint with A. In other words, A is the union of cells in $\mathcal{D}$.

THEOREM 13.17 (FINITE CELL DECOMPOSITION) *Given any semi-algebraic sets $A_1, \ldots, A_k$ of $\mathbb{R}^n$, there is a CAD of $\mathbb{R}^n$ partitioning each $A_1, \ldots, A_k$.* *QED*

EXAMPLE 13.18 Consider the semi-algebraic subset of $\mathbb{R}^2$ given by the formula

$$(x^2 + y^2 \leq 1) \vee \big((y = 0) \wedge (1 < x) \wedge (x < 2)\big) \vee \big((y \neq 0) \wedge (2 < x)\big).$$

This set is shown in part (a) of Fig. 13.8. In part (b) of this figure, a CAD of $\mathbb{R}^2$ is given, consisting of 25 cells, which partitions A. This CAD induces a CAD on the x-axis consisting of three (0)-cells and four (1)-cells (two of which are unbounded). On top of these intervals $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$-cells are built.

A proof of the finite cell decomposition theorem is given in (van den Dries, 1998, Ch. 3, Theorem 2.11). A key ingredient in this proof is the so-called *uniform finiteness property* of semi-algebraic sets. This property is also useful to obtain inexpressibility results, as will be shown in Sec. 5. To state this property, we need some definitions. A set A of $\mathbb{R}^{n+1}$ is called *finite over* $\mathbb{R}^n$ if for each $\vec{x} \in \mathbb{R}^n$ the fiber $A_{\vec{x}} = \{r \in \mathbb{R} \mid (\vec{x}, r) \in A\}$ is finite. We call A *uniformly finite over* $\mathbb{R}^n$ if there is an $N \in \mathbb{N}$ such that $|A_{\vec{x}}| \leqslant N$ for all $\vec{x} \in \mathbb{R}^n$. We then have:

THEOREM 13.19 (UNIFORM FINITENESS PROPERTY) *If $A \subseteq \mathbb{R}^{n+1}$ is a semi-algebraic set which is finite over $\mathbb{R}^n$, then A is uniformly finite over $\mathbb{R}^n$.*

As we will see in the next section, CAD is the basic tool for eliminating quantifiers. To be correct, we need an adaptation of the CAD to a given set of polynomials such that the sign of each of these polynomials is constant on each cell in the CAD. Moreover, in the context of quantifier elimination, CAD algorithms typically produce sample points in each cell, which enable to determine the sign of the polynomials. In Fig. 13.8, we have indicated a sample point for each cell in the CAD.

Triviality. We remark that until now, all results hold when we replace semi-algebraic by semi-linear. However, for the following result to be true one needs to work in the semi-algebraic setting.

EXAMPLE 13.20 Consider again the semi-algebraic set A and CAD of $\mathbb{R}^2$ of Example 13.18, shown in Fig. 13.8. Going from left to right, let $C_1 = (-\infty, a_1)$, $C_2 = \{a_1\}$, $C_3 = (a_1, a_2)$, $C_4 = \{a_2\}$, $C_5 = (a_2, a_3)$, $C_6 = \{a_3\}$ and $C_7 = (a_3, +\infty)$ be the (0) and (1)-cells on the x-axis. If one looks at the intersections of the cylinders $C_i \times \mathbb{R}$ with A, then it is clear that $A \cap (C_i \times \mathbb{R})$ is semi-algebraically homeomorphic to a product $C_i \times F_i$, where F_i is a semi-algebraic subset of $\mathbb{R}$. In this example, $F_1 = F_6 = \emptyset$, $F_2 = F_4 = F_5 = \{b_1\} \in \mathbb{R}$, $F_3 = [b_2, b_3] \subset \mathbb{R}$, and $F_7 = \mathbb{R} \setminus \{b_4\}$. In other words, the x-axis is decomposed into cells, such that A looks like a constant set above any two points in the same cell. One then says that the projection map $\pi : \mathbb{R}^2 \to \mathbb{R}$ on the x-axis is *trivial* over each of the cells $C_1, \ldots, C_7$.

We now formalize the intuition behind the example above. Let $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$ be two semi-algebraic sets and let $f : A \to B$ be a continuous semi-algebraic map. We can see A as a family of sets (i.e., fibers) $\{f^{-1}(\vec{b}) \mid \vec{b} \in B\}$. A semi-algebraic *trivialization of* f is a pair (F, λ) consisting of semi-algebraic set $F \subseteq \mathbb{R}^N$, for some N, and semi-algebraic map $\lambda : A \to F$ such that $(f, \lambda) : A \to B \times F$ is a homeomorphism.

Let $f : A \to B$ and suppose that f has a semi-algebraic trivialization. Then it is easy to show that all fibres are semi-algebraically homeomorphic to each other.

We call f *semi-algebraically trivial* if f has a semi-algebraic trivialization. Moreover, given $B' \subseteq B$, we say that f is *semi-algebraically trivial over* B' if the restriction of f to $f^{-1}(B')$ is semi-algebraically trivial.

THEOREM 13.21 (TRIVIALITY THEOREM) *Let $f : A \to B$ be a continuous semi-algebraic map as above. Then there is a finite partition of $B = B_1 \cup \cdots \cup B_\ell$ such that each B_i is semi-algebraic and f is semi-algebraically trivial over each B_i.*

3.3 The local conical structure of semi-algebraic sets

Let A be a semi-algebraic set of $\mathbb{R}^n$ and $\vec{p}$ a point of the closure of A. Let $B^n(\vec{p}, \varepsilon)$ be the closed ball with center $\vec{p}$ and radius ε and let $S^{n-1}(\vec{p}, \varepsilon)$ be the sphere with center $\vec{p}$ and radius ε.

We denote by $\mathrm{Cone}\big(\vec{p}, S^{n-1}(\vec{p}, \varepsilon) \cap A\big)$ the cone with vertex $\vec{p}$ and base $S^{n-1}(\vec{p}, \varepsilon) \cap A$, i.e., the set of points in $\mathbb{R}^n$ defined by $\lambda\vec{p} + (1 - \lambda)\vec{x}$ with $\lambda \in [0, 1]$ and $\vec{x} \in S^{n-1}(\vec{p}, \varepsilon) \cap A$. Let $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}$ denote the standard Euclidean norm.

THEOREM 13.22 (LOCAL CONICAL STRUCTURE) *For any point p of the closure of a semi-algebraic set A, there exists a number $\varepsilon > 0$ and a semi-algebraic homeomorphism*

$$h : B^n(\vec{p}, \varepsilon) \cap A \to \mathrm{Cone}\big(\vec{p}, S^{n-1}(\vec{p}, \varepsilon) \cap A\big)$$

such that $\|h(\vec{x}) - \vec{p}\| = \|\vec{x} - \vec{p}\|$ and $h\mid_{S^{n-1}(\vec{p},\varepsilon)\cap A} = \mathrm{Id}$.

A radius $\varepsilon > 0$ given by the previous theorem is called a cone radius of A in $\vec{p}$.

The local conical structure theorem is a direct consequence of the triviality theorem, for $f : A \to \mathbb{R}$ defined as $f(\vec{x}) = \|\vec{x} - \vec{p}\|$.

We examine the local conical structure of semi-algebraic sets of $\mathbb{R}^2$ in more detail. Consider the semi-algebraic set of $\mathbb{R}^2$ depicted in Fig. 13.9. For the point $\vec{p}$, we have indicated a cone radius ε of A in $\vec{p}$ by the dotted lines. The intersection $S^1(\vec{p}, \varepsilon) \cap A$ consists of a finite number of points and open intervals. If we denote open intervals that belong A by R^+ and intervals that belong to the complement of A by R^-, and if we similarly indicate points belonging to A by L^+ and points belonging to the complement by L^-, we can describe a the intersection $S^1(\vec{p}, \varepsilon) \cap A$ by means of a circular list over the alphabet $\{L^+, L^-, R^+, R^-\}$. We call this circular list the *cone type of A in $\vec{p}$*. The symbols L and R refer to lines and regions that arrive at p. There are two exceptions, however. For a point in the topological interior of A, we have that $S^1(\vec{p}, \varepsilon) = S^1(\vec{p}, \varepsilon) \cap A$, which we denote by F (for full). On the other hand, for an isolated point of A, we have $S^1(\vec{p}, \varepsilon) \cap A = \emptyset$, which we denote by E (for empty). For instance, the cone type of A in $\vec{p}$ in Fig. 13.9 is given by

$$(L^+R^-L^+R^-L^-R^+L^+R^-L^+R^-L^+R^+L^-R^+).$$

A semi-algebraic set A of $\mathbb{R}^2$ also has a local conical structure at infinity. To see this, we embed $\mathbb{R}^2$ as the (x, y)-plane in $\mathbb{R}^3$ and map A from this embedded plane onto the sphere $S^2((0, 0, 1), 1)$, that rests on the (x, y)-plane, in the direction of its north pole $(0, 0, 2)$. If we then add the north pole to this set as the point at infinity of the semi-algebraic set, rotate the sphere such that $(0, 0, 2)$

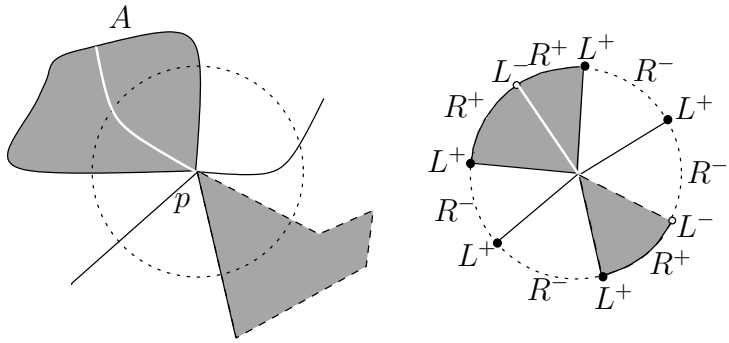


Figure 13.9. A semi-algebraic set A of $\mathbb{R}^2$ and the cone type of A in its points p given by the circular list $(L^+R^-L^+R^-L^-R^+L^+R^-L^+R^-L^+R^+L^-R^+)$.

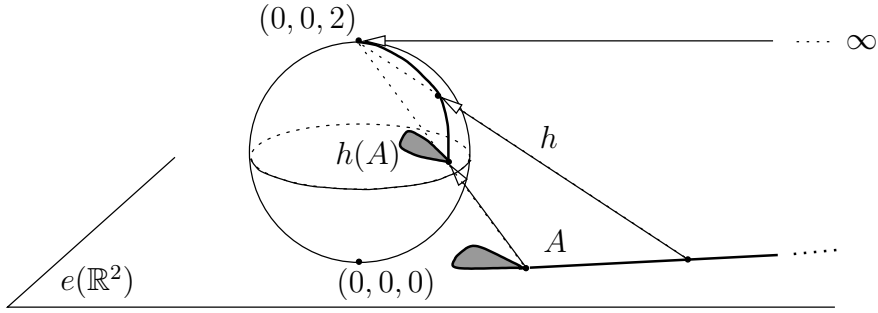


Figure 13.10. Illustration of the stereographical projection h.

becomes the origin, and stereographically project back on the xy-plane, then the local conical structure of $(0,0)$ in the resulting semi-algebraic set reveals the conical structure of the point at infinity in A.

This implies that there exists a $\varepsilon > 0$ such that $\{(x,y) \mid x^2+y^2 \geq \varepsilon^2\} \cap A$ is homeomorphic to $\{(\lambda x, \lambda y) \mid (x,y) \in S^1((0,0),\varepsilon) \cap A \wedge \lambda \geq 1\}$. We can indeed view the latter set as the cone with top ∞ and base $S^1((0,0),\varepsilon) \cap A$.

More formally, consider the embedding e of $\mathbb{R}^2$ in $\mathbb{R}^3$ that maps (x,y) to $(x,y,0)$. Let σ be the reflection of $\mathbb{R}^3$ defined by $(x,y,z) \mapsto (x,y,2-z)$. Finally, let $h : e(\mathbb{R}^2) \cup \{\infty\} \to S^2((0,0,1),1)$ be the homeomorphism of that maps the Alexandrov one-point compactification of $e(\mathbb{R}^2)$ stereographically onto the sphere $S^2((0,0,1),1)$, i.e., $h(x,y,0) = \frac{4}{4+x^2+y^2}(x,y,\frac{x^2+y^2}{2})$ and $h(\infty) = (0,0,2)$.

We define the the cone type of A in ∞ to be the cone type of the point $(0,0)$ in the set $e^{-1}(h^{-1}(\sigma(\{(0,0,2)\} \cup h(e(A))) \setminus \{\infty\}))$. We remark that the cone type of A in ∞ is (E) if and only if A is a bounded subset of $\mathbb{R}^2$.

Let A be a semi-algebraic set in $\mathbb{R}^2$ and let $\vec{p}$ be a point in the closure of A. Then it easily verified that for any two cone radii ε_1 and ε_2 of A in $\vec{p}$ (and ∞)

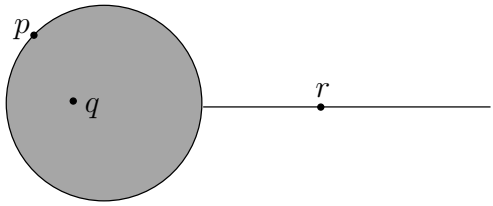


Figure 13.11. Types of regular points of a closed semi-algebraic set: $\Pi(p) = (R^- L^+ R^+ L^+)$, $\Pi(q) = F$ and $\Pi(r) = (R^- L^+ R^- L^+)$.

we get the same cone type. In other words, the notion of the cone type of A in $\vec{p}$ (and ∞) is well-defined.

Let $\mathcal{C}$ be the set of all possible cone types of semi-algebraic sets of $\mathbb{R}^2$. We define:

DEFINITION 13.23 Let A be a semi-algebraic set of $\mathbb{R}^2$. The *point-structure of* A is the function $\Pi(A)$ from $A \cup \{\infty\}$ to $\mathcal{C}$ that maps each point in the closure of A to its cone type.

An important observation is the following:

PROPOSITION 13.24 *The number of points in the closure of A with a cone different from $(R^- L^- R^+ L^-)$, $(R^- L^+ R^+ L^+)$, $(R^- L^+ R^- L^+)$, $(R^+ L^- R^+ L^-)$ and F is finite.*

We call point in the closure of A *singular* if it has a cone type different from $(R^- L^- R^+ L^-)$, $(R^- L^+ R^+ L^+)$, $(R^- L^+ R^- L^+)$, $(R^+ L^- R^+ L^-)$ or F. Otherwise, we say that a point in the closure of A is *regular*. It can be shown that only a finite number of cone types appear in A. Moreover, if A contains a point of cone type $(R^- L^- R^+ L^-)$, $(R^- L^+ R^+ L^+)$, $(R^- L^+ R^- L^+)$, $(R^+ L^- R^+ L^-)$ and F, then it must have infinitely many points of this cone type.

EXAMPLE 13.25 Suppose that A is a closed semi-algebraic set of $\mathbb{R}^2$. By the the observation above, there are infinitely many points in which A has one of the following three cone types $(R^- L^+ R^+ L^+)$, $(R^- L^+ R^- L^+)$, and (F). Fig. 13.11 illustrates these different cone types.

In Sec. 5, we will show the importance of the cone types and point structure for the expressibility of first-order logic over the reals.

3.4 Triangulations

An interesting question is whether semi-algebraic sets can exhibit more topological properties than semi-linear sets. The following results shows that this is not the case. Roughly speaking, the *triangulation theorem* states that each semi-algebraic set is homeomorphic to a semi-linear one. To make this more precise, we need the following notations.

Let $a_0, a_1, \ldots, a_k$ be $(k+1)$ affine independent points in $\mathbb{R}^n$. A *k-simplex* $(a_0, a_1, \ldots, a_k)$ is the set of points

$$(a_0, a_1, \ldots, a_k) = \left\{ \sum t_i a_i \mid \text{all } t_i > 0, \sum t_i = 1 \right\} \subseteq \mathbb{R}^n.$$

Note that k-simplex is of dimension k. Let σ be a k-simplex given by $(a_0, a_1, \ldots, a_k)$. The *closure of* σ, denoted by $\mathrm{cl}(\sigma)$ is the set of points

$$\mathrm{cl}(\sigma) = \left\{ \sum t_i a_i \mid \text{all } t_i \geqslant 0, \sum t_i = 1 \right\}.$$

A *face of* σ is a simplex corresponding to any nonempty subset of $(a_0, a_1, \ldots, a_k)$. A *complex in* $\mathbb{R}^n$ is a finite collection K of simplices in $\mathbb{R}^n$, such that for all $\sigma_1, \sigma_2 \in K$, either $\mathrm{cl}(\sigma_1) \cap \mathrm{cl}(\sigma_2) = \emptyset$, or $\mathrm{cl}(\sigma_1) \cap \mathrm{cl}(\sigma_2) = \mathrm{cl}(\tau)$, where τ is a common face of σ_1 and σ_2. We denote by $|K|$ the union of the simplices of K. From the definition it is clear that $|K|$ is a bounded semi-linear set of $\mathbb{R}^n$.

THEOREM 13.26 *Let A be a semi-algebraic set of $\mathbb{R}^n$. Then there exists a complex K in $\mathbb{R}^n$ and a semi-algebraic homeomorphism h such that $h(A) = |K|$, i.e., A is semi-algebraically homeomorphic to $|K|$.*

4. Query evaluation through quantifier elimination

In this section, we address in more detail how queries, expressible in the logics $\mathrm{FO}(+, <, 0, 1)$ and $\mathrm{FO}(+, \times, <, 0, 1)$, may be evaluated.

When we have a query expressed by a formula $\varphi(x_1, \ldots, x_m)$ over the vocabulary $(+, \times, <, 0, 1, S_1, \ldots, S_k)$ and we want to evaluate this query on a concrete input database over the schema $\mathcal{S} = (S_1, \ldots, S_k)$, given by quantifier-free formulas $\varphi_{S_1}(x_1, \ldots, x_{n_1}), \ldots, \varphi_{S_k}(x_1, \ldots, x_{n_k})$ in $\mathrm{FO}(+, \times, <, 0, 1)$ (n_i is the arity of S_i, $i = 1, \ldots, k$), we can proceed as follows:

- we plug-in the descriptions $\varphi_{S_1}(x_1, \ldots, x_{n_1}), \ldots, \varphi_{S_k}(x_1, \ldots, x_{n_k})$ of the input relations into the query formula $\varphi(x_1, \ldots, x_m)$ (this means that we replace each occurrence of some $S_i(v_1, \ldots, v_{n_i})$ in the query formula by $\varphi_{S_i}(v_1, \ldots, v_{n_i})$);
- this results in a formula over the vocabulary $(+, \times, <, 0, 1)$ that may contain quantifiers introduced by the query formula;
- next, we eliminate these quantifiers and obtain a quantifier-free description in $\mathrm{FO}(+, \times, <, 0, 1)$ of the output relation.

We remark that the same query evaluation strategy may be applied when $\times$ is omitted.

EXAMPLE 13.27 The formula

$$\exists \varepsilon (\varepsilon \neq 0 \wedge \forall x' \, \forall y' \, ((x - x')^2 + (y - y')^2 < \varepsilon^2 \rightarrow S(x', y')))$$

over the schema $(+, \times, <, 0, 1)$ expresses the topological interior of a set S in $\mathbb{R}^2$. When we want to evaluate the query expressed by this formula on the disk given by $x^2 + y^2 \leq 4$, we first replace $S(x', y')$ in the query formula by $(x')^2 + (y')^2 \leq 4$. This gives rise to the formula

$$\psi(x, y) = \exists\varepsilon(\varepsilon \neq 0 \wedge \forall x' \forall y' \, ((x-x')^2+(y-y')^2 < \varepsilon^2 \rightarrow (x')^2+(y')^2 \leq 4)).$$

The formula ψ contains three quantifiers. When we eliminate the quantifiers from ψ, we obtain as canonical quantifier-free description of the output the formula $x^2 + y^2 < 4$.

The reader might wonder why we bother about eliminating quantifiers. Indeed, simply plugging in the $\mathrm{FO}(+, \times, <, 0, 1)$-formulas of the input relations into the query formula yields a formula in $\mathrm{FO}(+, \times, <, 0, 1)$ that also describes the output. And these formulas, even though containing quantifiers, may be used in turn to describe an input to further queries. Even without eliminating quantifiers, we would have a formalism that has this *closure property*. Closure is a much desired property in database theory where it is considered important that outputs of queries may serve as input for further queries (compositionality of queries).

So, why is it so relevant to eliminate quantifiers? The answer lies in the question of what can we do with these formulas that describe output relations. Or rather, we should ask what we would like to do with these defining formulas.

Typical questions that are asked in database practice are the following:

- *Membership test*: for example, does $(1, 2)$ belong to the output relation given by $\varphi(u, v) = \exists x \exists y \, (u = x+y \wedge ((x = 1 \vee x = 2) \wedge y = 3)) \vee u = v$?
- *Emptiness test*: for example, is the set S given by $\exists x \exists y \, (z = x+y \wedge ((x = 1 \vee x = 2) \wedge y = 3))$ empty?

We observe that both questions, which are relevant to database practice, add up to deciding the truth of sentences of $\mathrm{FO}(+, \times, <, 0, 1)$. Indeed, to answer the membership test the truth of the sentence

$$\exists x \exists y \, (1 = x + y \wedge ((x = 1 \vee x = 2) \wedge y = 3)) \vee 1 = 2$$

has to be decided. For the second test, the truth of the sentence

$$\exists z \exists x \exists y \, (z = x + y \wedge ((x = 1 \vee x = 2) \wedge y = 3))$$

has to be determined.

Deciding the truth of sentences is possible in decidable theories like $(\mathbb{R}, +, <, 0, 1)$ and $(\mathbb{R}, +, \times, <, 0, 1)$. We discuss decision procedures for these theories in the subsections that follow.

4.1 Quantifier elimination for FO$(+, <, 0, 1)$

The theory of $(\mathbb{R}, +, <, 0, 1)$ has the following quantifier elimination property.

THEOREM 13.28 *The theory of* $(\mathbb{R}, +, <, 0, 1)$ *admits quantifier elimination. More specifically, this means that there is an algorithm that on input of a formula* $\varphi(x_1, \ldots, x_n)$ *over* $(+, <, 0, 1)$ *returns a quantifier-free formula* $\psi(x_1, \ldots, x_n)$ *that is equivalent to* $\varphi(x_1, \ldots, x_n)$ *over* $(\mathbb{R}, +, <, 0, 1)$.

We say that two formulas $\varphi(x_1, \ldots, x_n)$ and $\psi(x_1, \ldots, x_n)$ are *equivalent* if they define the same n-ary relation over $\mathbb{R}$.

For $(\mathbb{R}, +, <, 0, 1)$, there turns out to be a conceptually very simple quantifier elimination procedure that goes back to Fourier in 1826 and that was rediscovered by Motzkin in 1936 (Motzkin, 1936) and by several other researchers, even as late as the second half of the 20th century. We sketch the algorithm of Fourier now. Let us concentrate on the problem of eliminating a single existential quantifier from a formula $\varphi(x_1, \ldots, x_{m-1})$ of the form

$$\exists x_m \psi(x_1, \ldots, x_m),$$

where $\psi(x_1, \ldots, x_m)$ is a Boolean combination of atomic formulas of FO$(+, <, 0, 1)$. So, $\psi(x_1, \ldots, x_m)$ can be written as

$$\bigvee_{i=1}^{d} \bigwedge_{j=1}^{e_i} x_m \ \theta_{ij} \ c_{0ij} + \sum_{k=1}^{m-1} c_{kij} x_i$$

with $\theta_{ij} \in \{=, <, \leq, >, \geq\}$. If we abbreviate the terms $c_{0ij} + \sum_{k=1}^{m-1} c_{kij} x_i$ by t_{ij}, then we can remark that for any values given to the the variables $x_1, \ldots, x_{m-1}$, the terms t_{ij} can be ordered, let us say (after re-indexing) as $t_1 \leq t_2 \leq \cdots \leq t_k$. It is clear that for any two values of x_m taken strictly between some t_i and t_{i+1} (see Fig. 13.12), the truth value of $x_m \ \theta_{ij} \ c_{0ij} + \sum_{k=1}^{m-1} c_{kij} x_i$ is the same. The same is true if we take any two values of x_m strictly smaller than t_1 or strictly larger than t_k.

Therefore, when the $x_1, \ldots, x_{m-1}$ vary, the existential quantifier in $\exists x_m \psi$ $(x_1, \ldots, x_m)$ is equivalent expressible by the disjunction

$$\bigvee_{x_m = t_i \text{ or } x_m = 1/2(t_i + t_j) \text{ or } x_m = \pm\infty} \psi(x_1, \ldots, x_m).$$

In this disjunction, for x_m all values on and in between all of the t_{ij} are considered. In the above formula $x_m = \pm\infty$ can be achieved by considering all values $t_i \pm 1$ for x_m. It is clear that the above given disjunction suffices to replace the quantifier.

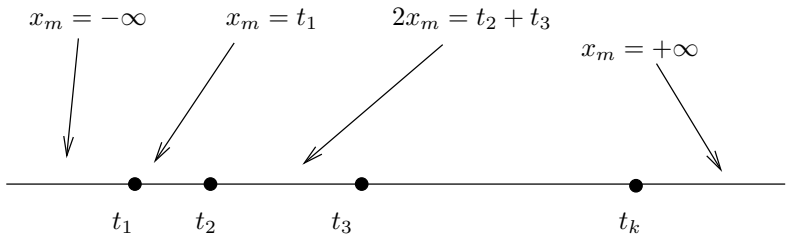


Figure 13.12. The relevant values for x_m.

We remark that this procedure takes exponential space in the size of the input formula and that the time complexity of this procedure is doubly exponential.

EXAMPLE 13.29 Suppose we want to eliminate the quantifier in

$$\varphi(x, y) = \exists z(x + z = y \wedge z < 2 + x).$$

To start with, we write this formula in the right form, namely $\exists z(z = y - x \wedge z < 2 + x)$. When we apply the procedure described above, we get the quantifier-free formula $\psi(x, y) =$

$$\begin{aligned}
&(y - x = y - x \wedge y - x < 2 + x) \vee \\
&\quad (2 + x = y - x \wedge 2 + x < 2 + x) \vee \\
&\quad (y - x - 1 = y - x \wedge y - x - 1 < 2 + x) \vee \\
&\quad (2 + x - 1 = y - x \wedge 2 + x - 1 < 2 + x) \vee \\
&\quad (y - x + 1 = y - x \wedge y - x + 1 < 2 + x) \vee \\
&\quad (2 + x + 1 = y - x \wedge 2 + x + 1 < 2 + x) \vee \\
&\qquad (y - x + 2 + x = 2(y - x) \wedge y - x + 2 + x < 2(2 + x)).
\end{aligned}$$

As mentioned above, we remark that the $-\infty$ and $+\infty$ from the algorithm are implemented by substracting and adding 1 from the terms $t_1 = y - x$ and $t_2 = 2 + x$ respectively. Also remark that many atomic formulas in this expression for $\psi(x, y)$ are trivially true or false. So, $\psi(x, y)$ could be further simplified.

4.2 Quantifier elimination for $\mathrm{FO}(+, \times, <, 0, 1)$

In the 1930s, Alfred Tarski showed that $\mathrm{FO}(+, \times, <, 0, 1)$ has the algorithmic quantifier elimination property too. Tarski published this result only in 1948 (Tarski, 1948).

THEOREM 13.30 *There is an algorithm that on input of a* $\mathrm{FO}(+, \times, <, 0, 1)$-*formula* $\varphi(x_1, \ldots, x_n)$ *returns a quantifier-free* $\mathrm{FO}(+, \times, <, 0, 1)$-*formula* ψ $(x_1, \ldots, x_n)$ *that is equivalent to the given formula* $\varphi(x_1, \ldots, x_n)$ *over* $(\mathbb{R}, +, \times, <, 0, 1)$.

The quantifier elimination procedure given by Tarski is based on a theorem by Sturm on real root counting and has a huge complexity (it is not elementary recursive), which makes it unsuitable for practical purposes.

EXAMPLE 13.31 A well-known example of quantifier elimination is the following. Consider the formula

$$\varphi(a, b, c) = a \neq 0 \wedge \exists x(ax^2 + bx + c = 0).$$

This formula describes triples (a, b, c) for which the quadratic equation $ax^2 + bx + c = 0$ has a real root. From high-school mathematics, we know that $\varphi(a, b, c)$ is equivalent to the quantifier-free formula

$$\psi(a, b, c) = a \neq 0 \wedge (b^2 - 4ac \geq 0).$$

Improvements to Tarski's procedure were proposed by Seidenberg, 1954, but a major breakthrough was achieved in Collins, 1975, which introduced the *cylindrical algebraic decomposition* (CAD) of semi-algebraic sets. His algorithm takes as input a system of polynomial equalities and inequalities that describe a semi-algebraic set in some $\mathbb{R}^n$. The algorithm returns a partitioning of $\mathbb{R}^n$ in a finite number of cells that are described by sign conditions on polynomials in n variables. These cells are actually accompanied by sample points in each of the cells that allow us to determine the sign conditions of these polynomials in these cells. The algorithm of Collins to compute a CAD has in the worst-case doubly-exponential sequential time complexity in the number of variables. It was the first quantifier elimination algorithm to be implemented, however. It has undergone numerous improvements, resulting in the implementation QEPCAD (Quantifier Elimination by Partial Cylindrical Algebraic Decomposition) by Hong, 1990. We refer to Caviness and Johnson, 1998 for a description at length of the current state of CAD.

A formal definition of a CAD was given in Sec. 3. It is beyond the scope of this chapter to give a full description of Collins' CAD algorithm, but we want to give an idea of the major steps in the algorithm.

Suppose the input of the CAD algorithm is an $\mathrm{FO}(+, \times, <, 0, 1)$-formula in prenex normal form

$$\varphi(u_1, \ldots, u_m) = \exists x_1 \cdots \exists x_n \bar{\varphi}(u_1, \ldots, u_m, x_1, \ldots, x_n).$$

Here, $\bar{\varphi}(u_1, \ldots, u_m, x_1, \ldots, x_n)$ is a Boolean combination of expressions of the form $p = 0$, $p > 0$, $p \geq 0$ or $p \neq 0$, where p is a polynomial. It is custom

in the quantifier elimination literature to distinguish between the variables $(x_1, \ldots, x_n)$ and parameters $(u_1, \ldots, u_m)$ of the given formula $\varphi(u_1, \ldots, u_m, x_1, \ldots, x_n)$. The goal is to eliminate the variables from the formula $\varphi(u_1, \ldots, u_m)$ via the computation of a CAD of $\bar{\varphi}(u_1, \ldots, u_m, x_1, \ldots, x_n)$.

The main construction steps in the construction of a CAD of the set $A = \{(u_1, \ldots, u_m, x_1, \ldots, x_n) \in \mathbb{R}^{m+n} \mid \bar{\varphi}(u_1, \ldots, u_m, x_1, \ldots, x_n)\}$ are:

- *the projection phase*: here the $(m + n)$-dimensional semi-algebraic set A is iteratively projected onto lower dimensional spaces ($\mathbb{R}^{m+n} \to \mathbb{R}^{m+n-1} \to \cdots \to \mathbb{R}^1$);
- *the basis phase*: here real roots are isolated in $\mathbb{R}^1$ and sample points are computed (using numeric methods);
- *the extension phase*: here, again in an iterative way ($\mathbb{R}^1 \to \mathbb{R}^2 \to \cdots \to \mathbb{R}^{m+n-1} \to \mathbb{R}^{m+n}$), a lifting to higher dimensions takes place. Stacks of cells (sections and sectors) are built, iteratively, together with sample points.

The output of the CAD algorithm is a sequence $C_1, \ldots, C_{m+n}$, where each C_i is a partition of $\mathbb{R}^i$ into cells. Each cell C is given by means of quantifier-free $\mathrm{FO}(+, \times, <, 0, 1)$-formula φ_C and a sample point. In particular, for each of the cells in the resulting decomposition of $\mathbb{R}^{m+n}$ it is recorded whether it belongs to the given semi-algebraic set A or not.

By construction, the set $\{(u_1, \ldots, u_m) \in \mathbb{R}^m \mid \varphi(u_1, \ldots, u_m)\}$ consists of all cells C in C_m for which there exists a cell C' in C_{m+n} which is in the stack $C \times \mathbb{R}^n$ of C and which belongs to A. Hence, a quantifier-free equivalent formula for $\varphi(u_1, \ldots, u_m)$ is obtained as a disjunction of all formulas φ_C describing cells C in C_m such that in the stack above C a cell of C_{m+n} belongs to A.

EXAMPLE 13.32 We illustrate the CAD algorithm using the three-dimensional set given by the quantifier-free $\mathrm{FO}(+, \times, <, 0, 1)$-formula

$$x^2 + y^2 + z^2 \leq 1 \vee (x^2 + y^2 + (z-2)^2 = 1 \wedge t \leq 5/2) \\ \vee (x^2 + y^2 + (z-3)^2 = 1 \wedge z > 5/2).$$

This set is depicted in Fig. 13.13.

In the projection phase this three-dimensional set is projected on (x, z)-plane and then this projected set is in turn projected on the z-axis (details omitted). On the real line certain "special points" are determined. In Fig. 13.14, these special points are coloured grey on the z-axis. These special points are always finite in number and they partition the line $\mathbb{R}$ into a finite number of points and open intervals (two of which are unbounded).

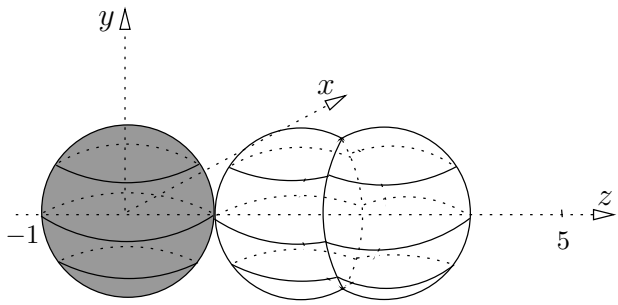


Figure 13.13. An example of a semi-algebraic set in $\mathbb{R}^3$.

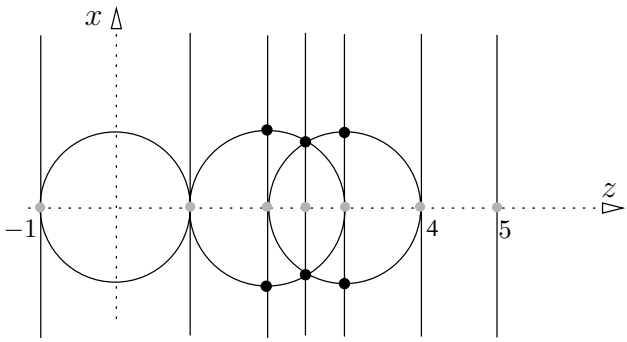


Figure 13.14. The 1- and 2-dimensional induced CADs of the semi-algebraic set of Fig. 13.13.

In the example of Fig. 13.13 and 13.14, there are 7 points and 8 intervals. This partition is called the one-dimensional induced CAD of the given set. Next, in the extension phase, stacks are built on the one-dimensional CAD. The stack above the second interval, for intance, consists of two curves (called sections) and three regions (called sectors), two of which are unbounded. The cells in these stacks form the two-dimensional induced CAD of the given set. Finally, stacks are built on these cells, resulting in the CAD of the given set, or to be more precise, of the description of the given set. For each of the cells in this decomposition of $\mathbb{R}^3$ it is recorded whether it belongs to the given semi-algebraic set or not.

During the 1990s, more efficient quantifier elimination algorithms were proposed. In 1990, Heintz, Roy, Solerno show a doubly exponential sequential time complexity in number of quantifier alternations, rather than in the number of quantifiers (Heintz et al., 1993). Later on, Heintz et al. show single exponential complexity if you work with alternative data structures, such as arithmetic Boolean circuits, to store systems of polynomial equalities and inequalities (Heintz et al., 1993). In the TERA project, the software *Kronecker* was developed and it is for the moment the most efficient software for quantifier

elimination (TERA-project, 1993) over the reals. The Kronecker implementation is decribed in (Giusti et al., 2001). We refer to (Basu et al., 2003b) for a detailed overview of algorithms in real algebraic geometry.

5. Expressiveness results

In this section, we discuss some results concerning the expressive power of $\mathrm{FO}(+, \times, <, 0, 1)$ and $\mathrm{FO}(+, <, 0, 1)$ as query languages for constraint databases.

The development of constraint databases has given rise to two directions of research. Firstly, classical relational database questions have been reconsidered. For instance, it is known that *graph connectivity* of finite relations is not expressible in first-order logic over relations (the same holds for other properties such as *parity*, *majority*, etc.). These expressiveness results can be re-addressed in the presence of arithmetical operations. Indeed, when we assume that the finite relations are embedded in the reals, we can ask whether connectivity, parity, majority, etc., are expressible when the vocabulary of first-order logic is extended with $+, \times, <, 0$ and 1. Secondly, expressiveness questions related to the possibility of representing infinite relations have been studied. In this section, we start by giving some results on finite relations and then show how they help to settle questions concerning the expressive power of $\mathrm{FO}(+, \times, <, 0, 1)$ on infinite relations.

5.1 Expressiveness results for finite databases

Here, we state a *generic collapse* result which allows to reduce—or collapse—expressiveness questions in the presence of arithmetic to the arithmetic-free case (Benedikt et al., 1996). We illustrate its implications on the first-order expressiveness of properties over finite databases over the reals. We also give the *dichotomy theorem* which gives a bound on the query result (in case it is finite) for first-order expressible queries (Benedikt and Libkin, 2000). This bound can be used to show inexpressibility results, as we shall illustrate.

Consider the following decision problems on finite relations:

- The decision problem MAJORITY for two finite sets S_1 and S_2 is: $\mathrm{MAJORITY}(S_1, S_2)$ is true if and only if $S_2 \subseteq S_1$ and $|S_1| \leq 2|S_2|$;

- The decision problem PARITY for a finite set S is: $\mathrm{PARITY}(S)$ is true if and only if $|S|$ is even.

The proof of the following lemma is a routine exercise in finite-model theory (Ebbinghaus et al., 1984). It can, e.g., be proven using the well-known technique of Ehrenfeucht-Fraïssé games. This lemma holds for arbitrary finite structures.

LEMMA 13.33 *On finite structures over the signature* $(<, S_1, S_2)$*, the decision problem* MAJORITY(S_1, S_2) *is not expressible in* FO$(<, S_1, S_2)$*. Likewise, on finite structures over the signature* $(<, S)$*, the decision problem* PARITY(S) *is not expressible in* FO$(<, S)$*.*

Benedikt, Dong, Libkin and Wong proved that any first-order formula over the reals that is invariant under monotone bijections from $\mathbb{R}$ to $\mathbb{R}$ is equivalently expressible on *finite* relations in the restriction of first-order logic that only uses order constraints (Benedikt et al., 1996). This collapse result was a breakthrough in the line of research towards understanding of the expressive power of first-order logic over the reals and related structures (Belegradek et al., 1996; Benedikt and Libkin, 1996; Benedikt and Libkin, 1997; Grumbach and Su, 1995; Grumbach et al., 1995; Paredaens et al., 1995; Stolboushkin and Taitslin, 1996).

Consider structures over the vocabulary $(+, \times, <, 0, 1, S_1, \ldots, S_k)$ that are expansions of $\mathbb{R}$ with k *finite* relations on $\mathbb{R}$. We call such structures *finite structures over the reals* (to emphasize the difference with finite structures in the sense of relational databases). A first-order formula over the vocabulary $(+, \times, <, 0, 1, S_1, \ldots, S_k)$ is called *order-generic* if on such structures, it is invariant under monotone bijections $f : \mathbb{R} \to \mathbb{R}$. Benedikt, Dong, Libkin, and Wong showed the following (Benedikt et al., 1996):

THEOREM 13.34 (COLLAPSE THEOREM) *For each order-generic formula in* FO$(+, \times, <, 0, 1, S_1, \ldots, S_k)$*, there exists a formula in* FO$(<, S_1, \ldots, S_k)$*, that is equivalent to it on finite structures over the reals. Furthermore, in the latter formula the quantifiers may be assumed to range only over the constants actually occurring in the relations* $S_1, \ldots, S_k$*.*

The following lemma, which specializes Lemma 13.33 from general finite ordered structures to finite structures over the reals, now follows directly from Lemma 13.33, Theorem 13.34 and the observation that the properties PARITY and MAJORITY of finite structures over the reals are invariant under monotone bijections from $\mathbb{R}$ to $\mathbb{R}$.

LEMMA 13.35 *On finite structures over the signatures* $(+, \times, <, 0, 1, S_1, S_2)$*, the decision problem* MAJORITY(S_1, S_2) *is not first-order expressible. Similarly, on finite structures over the signatures* $(+, \times, <, 0, 1, S)$*, the decision problem* PARITY(S) *is not first-order expressible.*

Apart from the above collapse results, there are also other results that are useful for showing that the expressive power of FO$(+, \times, <, 0, 1)$ is rather limited on finite structures over the reals.

THEOREM 13.36 (DICHOTOMY THEOREM) *Let* φ *be a* FO$(+, \times, <, 0, 1, S_1, \ldots, S_k)$*-formula. There exists a polynomial* p_φ *such that for any finite*

structure over the reals D, the query result of $\varphi(D)$ is either infinite or bounded by $p_\varphi(|D|)$, where $|D|$ denotes the number of elements in the structure D.

We provide an example of how to use the dichotomy theorem for showing inexpressibility results in the next section.

5.2 Inexpressibility results for infinite databases

We now apply the Dichotomy theorem to obtain an inexpressibility result for infinite databases. The example that we want to discuss concerns the linear ε-approximation of semi-algebraic sets.

DEFINITION 13.37 Let A be a semi-algebraic set of $\mathbb{R}^2$. A *(linear) ε-approximation* of A is a semi-linear set B of $\mathbb{R}^2$ which is homeomorphic to A via a homeomorphism $h : \mathbb{R}^2 \to \mathbb{R}^2$, and such that for any $\vec{p} \in A$, $d(\vec{p}, h(\vec{p})) < \varepsilon$.

THEOREM 13.38 *Let $\varepsilon > 0$ be a real number. There is no $, +, \times, <, 0, 1, S)$-formula that expresses a linear ε-approximation of a relation S in $\mathbb{R}^2$.*

Proof Consider the query Q that returns the empty set if the relation S does not consist of three non-collinear points, and otherwise returns the corner points of an ε-approximation of the circle determined by the three points of S. Here, a corner point is a point in which two straight-line segments make an angle different from 180 degrees.

Clearly, the construction of a circle through three points is expressible in $\mathrm{FO}(+, \times, <, 0, 1)$(the reader may want to verify this!). The same holds for the selection of the corner points of a semi-linear set (the reader may want to verify this as an exercise too). Hence, if we assume that the ε-approximation query can be expressed $\mathrm{FO}(+, \times, <, 0, 1, S)$, then Q is also expressible in $\mathrm{FO}(+, \times, <, 0, 1, S)$ by a formula φ. However, the number of corner points which is equal to $|\varphi(D)|$, can be made arbitrarily large by choosing D to consist of three far enough apart points. This contradicts the dichotomy theorem, which guarantees the existence of a polynomial p_φ such that the output of φ, when applied to D is bounded by $p_\varphi(|D|) = p_\varphi(3)$. QED

We remark that an alternative proof of Theorem 13.38 can be obtained using the uniform finiteness property of semi-algebraic sets instead of the dichotomy theorem.

5.3 Expressing topological properties of spatial databases

In this section, we will show that *topological connectivity* of planar geometric figures is not expressible in $\mathrm{FO}(+, \times, <, 0, 1)$. First, we remark that topological connectivity of a set in the plane is a query that is invariant under topological

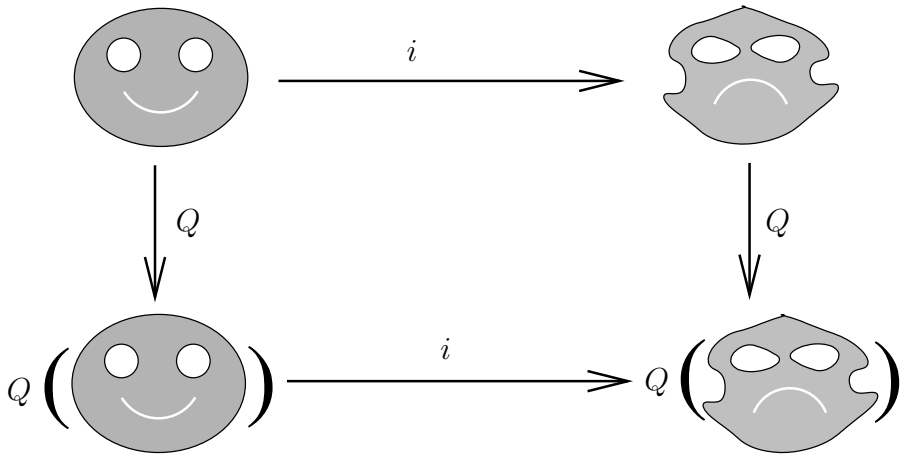


Figure 13.15. The query Q is topological.

transformations of the plane, i.e., it is invariant under isotopies $i : \mathbb{R}^2 \to \mathbb{R}^2$ (an isotopy is an orientation-preserving homeomorphism, i.e., it can be seen as a stretching transformation of the plane seen as a rubber sheet). Such queries are called *topological queries*. Formally, a query Q (over an input schema with one binary relation S) is called *topological* if for any two instances A and B of S for which there is an isotopy $i : \mathbb{R}^2 \to \mathbb{R}^2$ such that $i(A) = B$, $i(Q(A)) = Q(B)$ holds. This concept is illustrated in Fig. 13.15

We remark that deciding whether a query is topological is undecidable. For completeness we give the proof first presented in Paredaens et al., 1994.

THEOREM 13.39 *Testing whether a query expressible in* $\mathrm{FO}(+, \times, <, 0, 1)$ *is topological is undecidable.*

Proof Let $\mathcal{S}$ be the database schema consisting of a single unary relation S. For other schemas, the proof is similar. We will reduce the problem of deciding the truth of sentences of the $\forall^*$-fragment of number theory to the problem of deciding whether a query is topological. The $\forall^*$-fragment of number theory is known to be undecidable since Hilbert's 10th

We encode a natural number n by the unary finite relation $\mathrm{enc}(n) = \{0, 1, 2, \ldots, n\}$. A k-dimensional vector of natural numbers $(n_1, n_2, \ldots, n_k)$ is encoded by the relation

$$\mathrm{enc}(n_1, n_2, \ldots, n_k) = \mathrm{enc}(n_1) \cup (\mathrm{enc}(n_2) + n_1 + 2) \cup \\ \cdots \cup (\mathrm{enc}(n_k) + n_1 + 2 + \cdots + n_{k-1} + 2).$$

For a fixed k, the corresponding decoding is expressible in $\mathrm{FO}(+, \times, <, 0, 1)$ (the reader might want to verify this).

We now associate with each first-order sentence

$$\forall x_1 \forall x_2 \cdots \forall x_n \varphi(x_1, \ldots, x_n)$$

of number theory the following query Q_φ expressed by the $\mathrm{FO}(+, \times, <, 0, 1)$ formula over $\mathcal{S}$:

$$\begin{array}{rcl} Q_\varphi & = & \textbf{if } S \text{ encodes a vector } (n_1, \ldots, n_k) \in \mathbb{N}^k \textbf{ then} \\ & & \quad \textbf{if } \varphi(n_1, \ldots, n_k) \textbf{ then return } \emptyset \\ & & \quad \textbf{else return } 0 \\ & & \textbf{else return } \emptyset. \end{array}$$

It is easily verified that Q_φ is topological if and only if the sentence $\forall x_1 \forall x_2 \cdots \forall x_n \varphi(x_1, \ldots, x_n)$ is true. Therefore, if testing whether a query expressible in $\mathrm{FO}(+, \times, <, 0, 1)$ is topological would be decidable, so would be the $\forall^*$-fragment of number theory. QED

In the remainder of this section we investigate which databases can be distinguished by means of topological queries expressible in $\mathrm{FO}(+, \times, <, 0, 1, S)$. $\mathrm{FO}(+, \times, <, 0, 1, S)$. By definition, isotopic databases cannot be distinguished in such a way. It turns out that reverse direction does not hold. In general, it is not known when two databases are distinguishable by topological queries. There are two exceptions however. The first exception is for databases consisting of a single closed semi-algebraic set, the second exception is for databases consisting of, in general, many semi-algebraic sets, but in which only points of "regular" cone types are allowed. The latter case is discussed in (Grohe and Segoufin, 2002). We will only describe the case of closed databases. The following results are taken from (Kuijpers et al., 2000).

Cut- and glue transformations and the first-order inexpressibility of connectivity. Here we describe two transformations on closed semi-algebraic sets in $\mathbb{R}^2$. We call these transformations the *cut-* and the *glue transformation*. To apply the cut transformation to a set $A \subset \mathbb{R}^2$, one first needs to create locally (via a rubber-sheet transformation of the plane) a rectangular strip in A and then perform a cut as illustrated by the left to right direction in Fig. 13.16. So, this cut removes a rectangular part of the strip and perforates one of the remaining ends. The glue transformation is the inverse of the cut (illustrated by the right to left arrow in Fig. 13.16).

We will show that whenever two planar sets differ from each other by a *cut-* or *glue transformation*, they cannot be distinguished by a topological query expressed by a sentence in $\mathrm{FO}(+, \times, <, 0, 1, S)$.

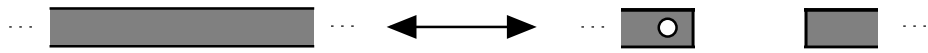


Figure 13.16. The cut and glue transformations on figures in $\mathbb{R}^2$.

THEOREM 13.40 *Let A and B be closed semi-algebraic sets in $\mathbb{R}^2$. If B is obtained from A by a cut- or glue transformation, then A and B are indistinguishable by a* $\mathrm{FO}(+, \times, <, 0, 1, S)$*-sentence that expresses a topological query.*

Proof Assume, for the sake of contradiction, that there exist closed semi-algebraic sets A and B that differ by one cut- or glue transformation but which can be distinguished by a first-order expressible topological sentence. Hence, there exists a first-order sentence φ, which expresses a topological query, such that $\varphi(A) = \text{TRUE}$ and $\varphi(B) = \text{FALSE}$.

Consider the decision problem MAJORITY about two finite sets of reals S_1 and S_2 (see above). We will prove the existence of a formula $\psi(x, y)$ in $\mathrm{FO}(+, \times, <, 0, 1, S_1, S_2)$ such that for any finite database $D = (D_1, D_2)$ over (S_1, S_2), we have that for $E_A(D) = \{(x, y) \in \mathbb{R}^2 \mid (\mathbb{R}, D) \models \psi(x, y)\}$, $\varphi(E_A(D)) = \text{TRUE}$ if and only if $\text{MAJORITY}(D_1, D_2)$ is FALSE.

By Lemma 13.35, this then yields the desired contradiction. This reduction technique is inspired by (Grumbach and Su, 1995).

Obviously, for any finite $D = (D_1, D_2)$, the part $D_1 \subseteq D_2$ can be tested in first-order logic. For given $D_1 = \{r_1, \ldots, r_n\}$ and $D_2 = \{a_1, \ldots, a_m\}$ with $0 < r_1 < \cdots < r_n$ and $0 < a_1 < \cdots < a_m$, we construct within the fixed rectangular part α of $\mathbb{R}^2$, where the cut-or-glue transformation takes place, a closed semi-algebraic set $E(D)$ consisting of interconnected strips.

This construction is similar to constructions in (Grumbach and Su, 1995) and is illustrated in Fig. 13.17 for $n = 6$ and $m = 4$. The construction is as follows. Take a rectangular subarea α' of α. Let (b_0, s_0) be the left bottom corner of α' and let h and w be its height and width. Then sets $D'_1 = \{s_0, \ldots, s_n\}$ and $D'_2 = \{b_0, b_1, \ldots, b_m, b_{m+1}, \ldots, b_{2m}\}$, with $s_i = s_0 + r_i h/r_n$ $(0 < i \leq n)$, $b_i = b_0 + a_i w/2a_m$ and $b_{m+i} = b_i + w/2$ $(0 < i \leq m)$ are constructed. Then, the following closed strips of $E(D)$ are constructed:

1. the filled convex quadrangle with corners (b_i, s_j), $((b_i + b_{i+1})/2, s_j)$, (b_{i+1}, s_{j+1}), $((b_{i+1} + b_{i+2})/2, s_{j+1})$ for $0 < i < 2m - 1$ and $0 \leq j < n$ and for $i = j = 0$,

2. the filled convex quadrangle with corners (b_{2m-1}, s_j), $((b_{2m-1} + b_{2m})/2, s_j)$, (b_{2m}, s_{j+1}), $(b_{2m}, (s_j + s_{j+1})/2)$ for $0 \leq j < n$,

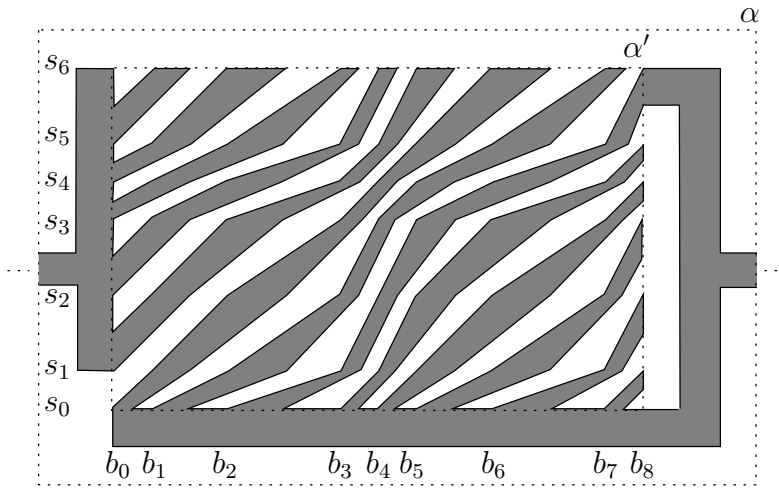


Figure 13.17. Construction of $E(D)$ for $D = (D_1, D_2)$ with $D_1 = \{1, 3, 5, 6, 7, 9\}$ and $D_2 = \{1, 3, 6, 7\}$ in the rectangular area α.

3. the filled convex quadrangle with corners $(b_0, (s_{j+1} + s_{j+2})/2)$, $((b_1 + b_2)/2, s_{j+2})$, (b_1, s_{j+2}), (b_0, s_{j+1}) for $0 \leq j < n - 1$.

Finally, a number of additional closed strips are added in the area $\alpha \setminus \alpha'$ (as illustrated in Fig. 13.17) to complete the construction of $E(D)$. Remark that the complete construction of $E(D)$, as described above, starting from D_1 and D_2 can be expressed for any $D = (D_1, D_2)$ by a formula $\mathrm{FO}(+, \times, <, 0, 1, S_1, S_2)$.

We then glue $E(D)$ to the part of A outside the cut-or-glue transformation area α and denote the resulting set by $E_A(D)$. Note that outside this area, $E_A(D)$ and B are identical.

Hence, there exists a formula ψ in $\mathrm{FO}(+, \times, <, 0, 1, S_1, S_2)$ such that for any D, we obtain a semi-algebraic set $E_A(D)$. Moreover, by construction $E(D)$ will be homeomorphic to the right part of Fig. 13.16 if $\textsc{majority}(D_1, D_2)$ is true, and homeomorphic to the left part of Fig. 13.16 otherwise.

Hence, in case of majority, $E_A(D)$ is homeomorphic to B, and in the other case it is homeomorphic to A. Since φ expresses a topological query which distinguishes between A and B, we can use φ to express MAJORITY. QED

With some more work additional cut and glue transformations can be proved to produce results that are indistinguishable from the original spatial figure. Three such transformations are shown in Fig. 13.18. In (a), we have a stip that can be cut (this time without producing a perforation in one of the sides). In

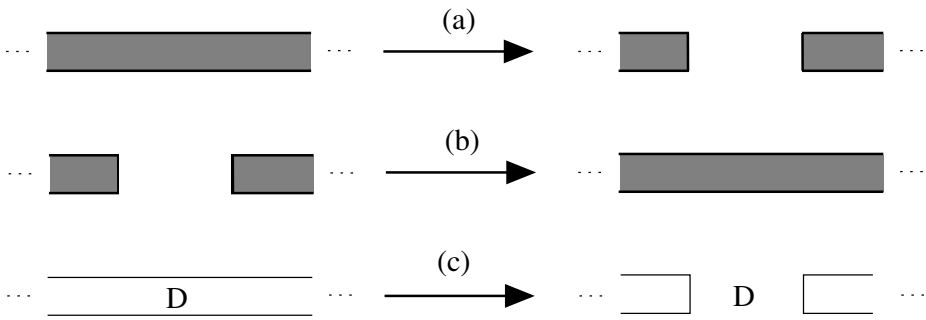


Figure 13.18. Three more cut and glue transformations on figures in $\mathbb{R}^2$.

Figure 13.19. One and two balls cannot be distinguished by a topological $\mathrm{FO}(+, \times, <, 0, 1)$-query.

(b), we have the reverse transformation of (a) and in (c) we see how parallel lines can be rewired, even when some data D is in between them.

REMARK 13.41 A direct consequence of Theorem 13.40 is that the connectivity query is not expressible in $\mathrm{FO}(+, \times, <, 0, 1)$. Indeed, this follows immediately from the observation that two disks can be transformed into a single disk using a cut and glue transformation (see Fig. 13.19).

We remark that Theorem 13.40 also holds when semi-algebraic is replaced by semi-linear. In this case, linear versions of the cut and glue transformations have to be considered.

The point structure of a closed semi-algebraic set in $\mathbb{R}^2$. In the previous section, we have shown that there are non-homeomorphic sets that nevertheless are indistinguishable by $\mathrm{FO}(+, \times, <, 0, 1)$-sentences expressing topological properties. More specifically, the cut- or glue transformations relate some indistinguishable databases. However, there are a few other transformations with the same property and the question now arises as to which databases can

be transformed into each other using one of these transformations. For closed databases, we can characterize this exactly.

First, recall the definition 13.23 of the point structure of a database.

DEFINITION 13.42 Let A and B be closed semi-algebraic sets in $\mathbb{R}^2$. We say that $\Pi(A)$ is *isomorphic* to $\Pi(B)$ (denoted by $\Pi(A) \cong \Pi(B)$) if there is a bijection f from $A \cup \{\infty\}$ to $B \cup \{\infty\}$ with $f(\infty) = \infty$, such that $\Pi(A) = \Pi(B) \circ f$.

The main result in this context is that indistinguishability by topological queries can be expressed in terms of point-structure isomorphism.

THEOREM 13.43 *Let A and B be closed semi-algebraic sets in $\mathbb{R}^2$. The sets A and B are indistinguishable by topological* $\mathrm{FO}(+, \times, <, 0, 1)$*-queries if and only if* $\Pi(A) \cong \Pi(B)$.

Proof We briefly sketch the proof of this theorem. If two closed semi-algebraic sets have a different point structure than they can be distinguished by a topological query that expresses that the one set contains a different number of points than the other set with a specific cone. In Sec. 5.4, we will discuss in more detail how a cone of a point can be expressed in $\mathrm{FO}(+, \times, <, 0, 1)$.

If two closed semi-algebraic sets in $\mathbb{R}^2$ have the same point structure, they can be transformed into the same canonical semi-algebraic set using the cut and glue transformations shown in Fig. 13.18. First, two-dimensional lobes are cut around singular points, then the lines are cut into lobes. This produces "flowers" that may be connected by stems. This rewriting process is illustrated in Fig. 13.20. QED

One may wonder whether this characterization holds for arbitrary semi-algebraic sets in $\mathbb{R}^2$ too. This is not the case as it can be shown that the two sets shown in Fig. 13.21 can be distinguished by a topological $\mathrm{FO}(+, \times, <, 0, 1)$-sentence even though they have isomorphic cone structures (Grohe and Segoufin, 2002).

Theorem 13.43 gives us an idea of which closed semi-algebraic sets in the plane are distinguishable by topological first-order queries, but it doesn't give us a full picture of the expressive power of the topological fragment of $\mathrm{FO}(+, \times, <, 0, 1)$. There are results that characterize this expressive power, however. It has been shown that the topological fragment of $\mathrm{FO}(+, \times, <, 0, 1)$ just allows us to formulate queries that talk about types of cones appearing in a semi-algebraic sets and on the number of points having particular cones (Benedikt et al., 2006).

5.4 Expressing the cone radius

As we have seen, the local conical property of semi-algebraic sets plays a prominent role in the study of first-order expressiveness of topological

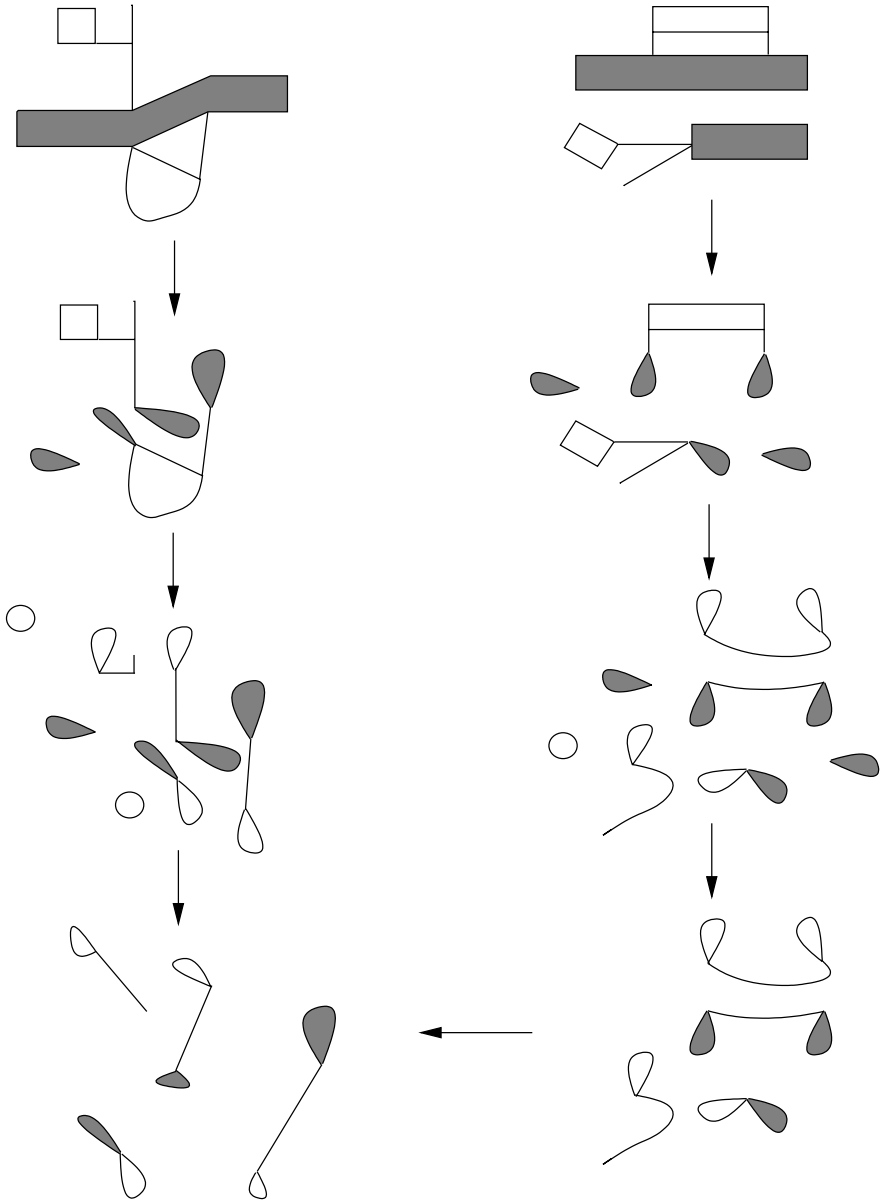


Figure 13.20. The transformation of two closed semi-algebaric sets (on the left and right hand top) into their canonical form (left bottom).

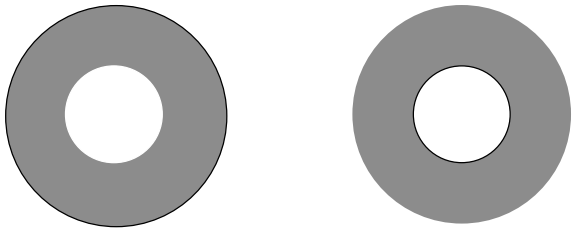


Figure 13.21. Semi-open annuli with the opposite open sides can be distinguished by a topological $\mathrm{FO}(+, \times, <, 0, 1)$-query.

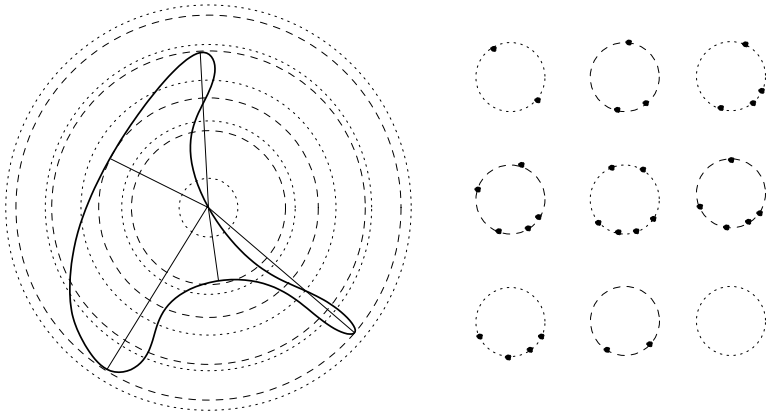


Figure 13.22. Illustration of how the intersections $S^1(\vec{p}, r) \cap A$ change in terms of r.

properties. In this section, we will show that $\mathrm{FO}(+, \times, <, 0, 1)$ is expressive enough to find a cone radius for each point in the database.

More specifically, the following result holds (Geerts, 2003).

THEOREM 13.44 *There exists an* $\mathrm{FO}(+, \times, <, 0, 1, S_1, S_2)$*-formula* $\varphi(r)$*, with* S_1 *and* S_2 *of arity* n*, that for a semi-algebraic set* A *in* $\mathbb{R}^n$ *and a point* $\vec{p}$ *in* $\mathbb{R}^n$*, defines one* r *in* $\mathbb{R}$ *for which* $(A, \{\vec{p}\}) \models \varphi(r)$ *and which is a cone radius for* A *in* $\vec{p}$*.*

To give the complete proof of this theorem would lead us too far afield. Instead, we provide the intuition behind the proof. Moreover, we only consider the case when $n = 2$ and assume that A is a semi-algebraic set consisting only of points of cone type (LL).

Consider the semi-algebraic set depicted on the left of Fig. 13.22. On the right of this picture, we have shown the intersections of circles of various radii, centered around $\vec{p}$ with A. As can be seen, each time there exists a point $\vec{q}$

in A which has a tangent line perpendicular to the line going through $\vec{p}$ and $\vec{q}$ (depicted by the dashed circles), the topological type of the intersection, which in this case is nothing else than the number of points, changes.

This observation can be formalized using a variant of the triviality theorem which states that the topological type of $S^1(\vec{p}, r_1) \cap A$ and $S^1(\vec{p}, r_2) \cap A$ are the same (meaning that there exists an homeomorphism between these two sets) if for any $r \in [r_1, r_2]$ there exists no point $\vec{q}$ such that $d(\vec{p}, \vec{q}) = r$ and the tangent line in $\vec{q}$ is perpendicular to $\vec{q} - \vec{p}$.

It can be shown that there exists an FO$(+, \times, <, 0, 1, S_1, S_2)$-formula φ which returns for any pair $(A, \vec{p})$ all point $\vec{q}$ which have a tangent line perpendicular to the vector $\vec{q} - \vec{p}$. Using φ we define

$$r_{\vec{p}} = \frac{1}{2} \min\{d(\vec{p}, \vec{q}) \mid (A, \vec{p}) \models \varphi(\vec{q})\}.$$

It is clear that for any $(A, \vec{p})$, $r_{\vec{p}}$ is expressible by a first-order formula. By definition, there is no point $\vec{q}$ with a tangent line perpendicular to $\vec{q} - \vec{p}$ for any $0 < r \leq r_{\vec{p}}$. In other words, for all $0 < r \leq r_{\vec{p}}$, all interesections $S^1(\vec{p}, r) \cap A$ are homeomorphic, and it can be shown that $r_{\vec{p}}$ is indeed a cone radius of A in $\vec{p}$.

We briefly mention that for the general case, we have to decompose A first into parts on which a tangent space can be defined. Moreover, this has to be shown to be first-order expressible. Next, similar to the case above, a cone radius of a point $\vec{p}$ can then be defined as a radius smaller than any point $\vec{q}$ with a tangent space perpendicular to $\vec{q} - \vec{p}$, where the tangent space is taken relative to the decomposition of A.

5.5 The expressiveness of FO$(+, \times, <, 0, 1)$ and FO$(+, <, 0, 1)$

We end this section with a remark on the comparison of the expressive powers of FO$(+, \times, <, 0, 1)$ and of FO$(+, <, 0, 1)$. We will illustrate that the expressive power of FO$(+, <, 0, 1)$ is less than that of FO$(+, \times, <, 0, 1)$.

The topological interior of a two-dimensional set S can be expressed in FO$(+, \times, <, 0, 1)$ by the formula

$$\exists \varepsilon(\varepsilon \neq 0 \wedge \forall x' \forall y'((x - x')^2 + (y - y')^2 < \varepsilon^2 \rightarrow S(x', y')).$$

Since the topology of $\mathbb{R}^2$ based on open discs is equivalent to the one based on open rectangles, we can equivalently express the topological interior of a semi-algebraic subset of $\mathbb{R}^2$ in FO$(+, <, 0, 1)$ by the formula

$$\exists \varepsilon(\varepsilon > 0 \wedge \forall x' \forall y'((\mid x - x' \mid < \varepsilon \wedge \mid y - y' \mid < \varepsilon) \rightarrow S(x', y')).$$

But there are other queries for which the multiplication seems to be really necessary to express them in first-order logic. If we want to express that a

two-dimensional semi-linear set is *convex*, for instance, then we can do this in $\mathrm{FO}(+, \times, <, 0, 1)$ with the formula

$$\forall \vec{x} \forall \vec{y} (S(\vec{x}) \wedge S(\vec{y}) \rightarrow \forall \lambda (0 \leq \lambda \leq 1 \rightarrow S(\lambda \vec{x} + (1 - \lambda)\vec{y})).$$

Clearly, in the subexpression $\lambda \vec{x} + (1 - \lambda)\vec{y}$ multiplication is used and it may seem difficult to imagine that convexity of semi-linear sets might be expressible without multiplication.

But it turns out that we have the following property (Vandeurzen et al., 1996).

PROPOSITION 13.45 *A semi-linear set of $\mathbb{R}^n$ is convex if and only if it is closed under taking midpoints.*

We remark this property does not hold for subsets of $\mathbb{R}^n$ that are not semi-linear, e.g., think of $\mathbb{Q}^n$.

We can therefore express convexity of semi-linear sets by the $\mathrm{FO}(+, <, 0, 1)$-formula

$$\forall \vec{x} \forall \vec{y} (S(\vec{x}) \wedge S(\vec{y}) \rightarrow \exists \vec{z} (2\vec{z} = \vec{x} + \vec{y} \wedge S(\vec{z})).$$

It is not true, however, that all queries expressible in $\mathrm{FO}(+, \times, <, 0, 1)$ are also expressible in $\mathrm{FO}(+, <, 0, 1)$. Indeed, $\mathrm{FO}(+, \times, <, 0, 1)$ is clearly more expressive than $\mathrm{FO}(+, <, 0, 1)$ as far as queries that return some n-dimensional result are concerned. For example, the constant query that returns on any input the n-dimensional unit sphere, for instance, is not expressible in $\mathrm{FO}(+, <, 0, 1)$. But what about Boolean queries? It turns out that $\mathrm{FO}(+, \times, <, 0, 1)$ is again more expressive than $\mathrm{FO}(+, <, 0, 1)$ as far as Boolean queries are concerned, as is illustrated by the following theorems. The first theorem is from (Afrati et al., 1994) and was proved using Ehrenfeucht-Fraïssé games.

THEOREM 13.46 *The Boolean query deciding whether a semi-linear set S contains real numbers u and v satisfying $u^2 + v^2 = 1$ is expressible in* $\mathrm{FO}(+, \times, <, 0, 1, S)$*, but not in* $\mathrm{FO}(+, <, 0, 1, S)$.

Another example of a property that is not expressible in $\mathrm{FO}(+, <, 0, 1)$ is from (Benedikt and Keisler, 2000).

THEOREM 13.47 *The Boolean query deciding whether a semi-linear subset S of $\mathbb{R}^2$ contains a line is expressible in* $\mathrm{FO}(+, \times, <, 0, 1, S)$*, but not in* $\mathrm{FO}(+, <, 0, 1, S)$.

6. Extensions of logical query languages

In this section we will extend first-order logic with operators in order to increase its expressive power. We begin with introducing transitive closure logics (Geerts and Kuijpers, 2000; Geerts and Kuijpers, 2005) and (Kreutzer, 2001). Next, we extend first-order logic with a while-loop (Gyssens et al., 1999). We conclude the section by extending first-order logic with specific operators for topological properties (Benedikt et al., 2003).

6.1 First-order logic with all-purpose operators

Let us first make a small digression to the relational database setting. Here, a database D is a finite model of the signature $\mathcal{S}$, where $\mathcal{S}$ consists of a binary relation name S. The transitive closure of D is usually computed in stages X_i, $i = 0, 1, 2, \ldots$. Initially, $X_0 = D$ and for $n \geqslant 0$ we have

$$X_{n+1} = X_n \cup \big\{(x, y) \mid \exists z\big(X_n(x, z) \wedge D(z, y)\big)\big\}.$$

Since the number of pairs in D is finite and by construction, $X_n \subseteq X_{n+1}$, after finitely many steps we end up with the situation that $X_{n+1} = X_n$. The fixed-point X_n is then the transitive closure of D.

It is well-known that the transitive closure of D cannot be computed by a query expressible in first-order logic over $\mathcal{S}$. In order to be able to express the transitive closure of D one has studied the extension of first-order logic with transitive closure (Ebbinghaus and Flum, 1995).

EXAMPLE 13.48 We show that transitive closure logics can express that a graph $G = (V, E)$ is connected. Indeed, let D be the (binary) database containing the edge relation E of G. If we interpret the formula

$$\mathrm{TC}_{x;y} S(x, y)$$

as an expression which evaluates on D to the transitive closure of D, then the expression

$$\forall s \forall t \big[\mathrm{TC}_{x;y} S(x, y)\big](s, t)$$

evaluates to true on D if and only if the corresponding graph G is a connected, i.e., if for any two s, t in the domain of D, i.e., the set of vertices V of G, there exists a sequence $(v_0, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n)$ such that $D(v_i, v_{i+1})$ for $i = 0, 1, \ldots, n-1$ and $s = v_0$ and $t = v_n$.

Motivated by this relational example, we start by adding the transitive closure operator to first-order logic in the constraint database setting. The resulting query language will be denoted by $\mathrm{FO}(+, \times, <, 0, 1)$+TC.

First-order logic with transitive-closure operator. A formula in $\mathrm{FO}(+, \times, <, 0, 1)$+TC is a formula built in the same way as an $\mathrm{FO}(+, \times, <, 0, 1)$ formula, but with the following extra formation rule: if $\psi(\vec{x}, \vec{y})$ is a formula with $\vec{x}$ and $\vec{y}$ k-tuples of real variables, with all free variables of ψ among $\vec{x}$, $\vec{y}$, and if $\vec{s}$, $\vec{t}$ are k-tuples of real variables, then

$$\big[\mathrm{TC}_{\vec{x};\vec{y}}\, \psi(\vec{x}, \vec{y})\big](\vec{s}, \vec{t}) \tag{13.1}$$

is also a formula which has as free variables those in $\vec{s}$ and $\vec{t}$.

We will distinguish between $\mathrm{FO}(+,<,0,1)+\mathrm{TC}$ and $\mathrm{FO}(+,\times,<,0,1)+\mathrm{TC}$ depending on whether only linear (i.e., $\times$ is not allowed) or also non-linear constraints are allowed.

To explain the semantics of a subformula of the above form (13.1), we compute again the following stages

$$X_0 = \psi(D), \qquad X_{n+1} = X_n \cup \big\{(\vec{x},\vec{y}) \mid \exists \vec{u}\big(X_n(\vec{x},\vec{u}) \wedge X_0(\vec{u},\vec{y})\big)\big\},$$

until the fixed-point, which we denote by X_∞, is reached. Then the semantics of $\big[\mathrm{TC}_{\vec{x};\vec{y}}\,\psi(\vec{x},\vec{y})\big](\vec{s},\vec{t})$ is defined as $(\vec{s},\vec{t})$ belonging to the $2k$-ary relation X_∞.

The question is now how a formula in $\mathrm{FO}(+,\times,<,0,1)+\mathrm{TC}$ is evaluated. Assume that ψ is an $\mathrm{FO}(+,\times,<,0,1)$-formula. Then we simply can compute the quantifier-free description of X_{n+1} recursively by evaluating the corresponding $\mathrm{FO}(+,\times,<,0,1)$-expressions. After each computation we then test whether $X_n = X_{n+1}$. If this holds, we have obtained the fixed-point and test whether $(\vec{s},\vec{t})$ is in X_n.

In general, the semantics of a formula φ in $\mathrm{FO}(+,\times,<,0,1)+\mathrm{TC}$ is evaluated in the standard bottom-up fashion. The result of the evaluation of subformulas is passed on to formulas that are higher up in the parsing tree of φ.

However, in this context, we face the well-known fact that recursion involving arithmetic over an unbounded domain, such as the polynomial inequalities over the reals in our setting, is no longer guaranteed to terminate. In other words, the computation of X_∞ might not terminate. Hence, the property of effective computability of queries expressible in $\mathrm{FO}(+,\times,<,0,1)$ or $\mathrm{FO}(+,<,0,1)$ is lost when extending these logics with recursion. In case the computation of X_∞ does not terminate, then the semantics of the formula (13.1) (and any other formula in which it occurs as subformula) is undefined.

We therefore have to distinguish between formulas for which the evaluation terminates, we call such formulas *terminating*, and formulas for which the evaluation does not terminate, i.e., the *non-terminating* formulas. To illustrate this difference, we now provide an example of a terminating and non-terminating formula in $\mathrm{FO}(+,\times,<,0,1)+\mathrm{TC}$.

EXAMPLE 13.49 Let $\mathcal{S}$ consist of the binary relation S. Consider the $\mathrm{FO}(+,\times,<,0,1)+\mathrm{TC}$ formula

$$\big[TC_{x;y}S(x,y)\big](s,t)$$

and let $D = \{(x,y) \mid y = 2x\}$. We have for each $n \geqslant 0$,

$$X_n = \big\{(x,y) \mid \exists i \in \mathbb{N}, i \leqslant n, y = 2^i x\big\}.$$

It is clear that for all $n \geqslant 0$, $X_n \neq X_{n+1}$ and hence we need to compute infinitely many stages until the fixed point X_∞ is reached. We have depicted this example in Fig. 13.23. It easy to see t hat X_∞ is not a semi-algebraic set.

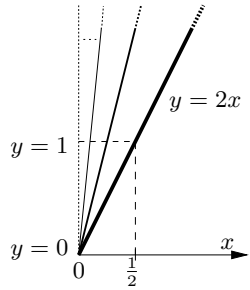


Figure 13.23. Computation of $\big[TC_{x;y}S(x,y)\big](s,t)$ for $D = \{(x,y) \mid y = 2x\}$ (thickest line). Consecutive stages are drawn in decreasingly finer lines.

On the other hand, even when X_∞ is semi-algebraic, its computation may still be non-terminating. Moreover, one may wonder whether the non-termination of the $\mathrm{FO}(+,\times,<,0,1)$+TC formula in Example 13.49 is caused by the unboundedness of the input database. However, it easy to see that even bounded input databases may cause non-termination.

Since $\mathrm{FO}(+,\times,<,0,1)$ is a sub-language of $\mathrm{FO}(+,\times,<,0,1)$+TC, there are infinitely many terminating formulas. We now give an example of a terminating formula which is not in $\mathrm{FO}(+,\times,<,0,1)$.

EXAMPLE 13.50 Let the database schema $\mathcal{S}$ consist of the unary relation name S. Consider the formula

$$\big[TC_{x;y}\varphi(r,x,y) \wedge S(r)\big](s,t)$$

where $\varphi(r,x,y)$ defines the graph of the continuous piecewise affine function that maps x to

$$y = \begin{cases} 0 & \text{if } 0 \leq x \leq \frac{1}{r}, \\ x - \frac{1}{r} & \text{if } \frac{1}{r} < x < 1, \\ 1 - \frac{1}{r} & \text{if } x = 1. \end{cases}$$

Then, for $D = \{p\}$ and $p \in \mathbb{N}$ we have for each $n \geqslant 0$,

$$X_n = \bigcup_{i=1}^{n} \big\{(x,y) \mid \varphi(\frac{n}{p}, x, y)\big\}.$$

It is clear that $X_{p+1} = X_p$ and hence this $\mathrm{FO}(+,\times,<,0,1)$+TC formula is terminating. Moreover, it expresses the query Q_{nat} which tests whether the number stored in a database is a natural number. We have illustrated this example in Fig. 13.24. It is easily verified that the query Q_{nat} cannot be expressed in the logic $\mathrm{FO}(+,\times,<,0,1)$.

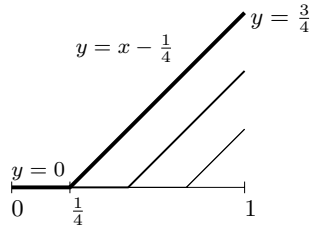


Figure 13.24. Example of the terminating $\mathrm{FO}(+, \times, <, 0, 1)$+TC formula of Example 13.50 for $D = \{4\}$.

As a result of the above example and Example 13.50, we have that $\mathrm{FO}(+, \times, <, 0, 1)$ is strictly less expressive than $\mathrm{FO}(+, \times, <, 0, 1)$+TC.

As explained in Example 13.48, transitive closure logic can express the connectivity of finite graphs. Similarly, $\mathrm{FO}(+, \times, <, 0, 1)$+TC can express the connectivity of constraint databases. We first consider the case of linear constraints.

Let $\mathcal{S}$ be a schema with one relation name S of arity n. Consider the following $\mathrm{FO}(+, \times, <, 0, 1)$+TC formula $\texttt{connected}(S)$:

$$\forall \vec{s}\, \forall \vec{t} \big(\big(S(\vec{s}) \wedge S(\vec{t}) \big) \rightarrow [\mathrm{TC}_{\vec{x};\vec{y}}\, \texttt{lineconn}](\vec{s}, \vec{t}) \big) ,$$

where $\texttt{lineconn}(\vec{x}, \vec{y})$ is the formula

$$\forall \lambda (0 \leq \lambda \leq 1 \wedge \forall \vec{t} (\vec{t} = \lambda \vec{x} + (1 - \lambda) \vec{y} \rightarrow S(\vec{t})).$$

We now claim that a pair of points $(\vec{p}, \vec{q})$ belongs to transitive closure of $\texttt{lineconn}(D)$ (with D semi-linear) if and only if $\vec{p}$ and $\vec{q}$ belong to the same connected component. Hence, if we can show that `connected` is a terminating formula, then this implies that `connected` expresses connectivity of semi-linear sets.

THEOREM 13.51 *The formula* `connected` *terminates on all linear constraint databases D over $\mathcal{S}$ and correctly expresses connectivity of semi-linear sets.*

Proof Since D is semi-linear, two points $\vec{p}$ and $\vec{q}$ belong to the same connected component of D if and only if there exists a piecewise linear path from $\vec{p}$ to $\vec{q}$ lying entirely in D. This follows directly from the semi-linear version of Theorem 13.14 and Remark 13.15. So, indeed $[\mathrm{TC}_{\vec{x};\vec{y}}\, \texttt{lineconn}](\vec{s}, \vec{t})$ holds if and only if $\vec{s}$ and $\vec{t}$ belong to the same connected component of D.

To conclude that the evaluation of the transitive closure in the formula `connected` ends in finitely many steps, we need to show that there exists an upper bound on the number of line segments in S, needed to connect any two points in the same connected component of S. Now, any semi-linear set can be decomposed into a finite number of convex sets (van den Dries, 1998,

Ch. 8, Exercise 2.14 (2)). The finiteness of this decomposition yields the desired bound since any two points in a convex set are connected by a single straight line segment. We have illustrated this in Fig. 13.25 for $n = 2$. QED

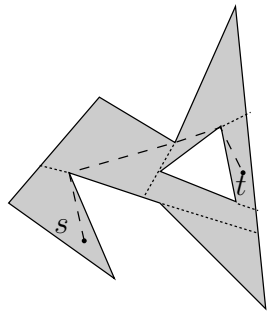


Figure 13.25. A semi-linear set decomposed in convex sets. The boundaries of the convex sets are shown in dotted lines. We have depicted a piecewise linear path (dashed line) between points s and t.

The above $\mathrm{FO}(+, \times, <, 0, 1)$+TC-formula `connected` cannot be applied to arbitrary semi-algebraic inputs, while still guaranteeing termination. Indeed, the evaluation of `connected` on the binary relation depicted in Fig. 13.26, consisting of the points lying strictly between the parabola $y = x^2$ and the translated one $y = x^2 + 1/2$, will not terminate. Although any two points in this set can be connected by a finite number of line segments, there is no upper bound on the number of segments needed to connect two points.

There are also examples of semi-algebraic sets s for which there exist two points in a connected component that cannot be connected by any finite piecewise linear path. Here we take as example the set defined by $(y^3 - x^2 \geq 0 \wedge x <$

Figure 13.26. A semi-algebraic set, the points lying strictly between the parabola $y = x^2$ and the translated parabola $y = x^2 + 1/2$, on which termination is not satisfied.

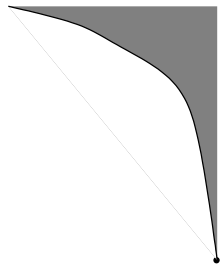


Figure 13.27. A semi-algebraic set with a cusp point.

$0 \wedge y < 1) \vee (x = 0 \wedge y = 0)$, depicted in Fig. 13.27. The cusp point $(0, 0)$ at the bottom cannot be connected by a line segment to any point in the interior of this set. So, the query expressed by `connected` terminates after two iterations, but it does not contain all pairs of points that are in the connected component. This is obviously because piecewise-linear connectivity does not correspond to connectivity for arbitrary semi-algebraic sets.

Basically, Fig. 13.26 and 13.27 illustrate the only two cases where the `connected` query does not work correctly (i.e., cusp points and cusp points towards ∞).

Nevertheless, we can also express connectivity of arbitrary semi-algebraic sets. The proof of this results is a reduction to the linear case. This reduction is possible by the following result.

THEOREM 13.52 *There exists a terminating* $\mathrm{FO}(+, \times, <, 0, 1)$+TC *formula which expresses the linearization query "Return a semi-linear set which is homeomorphic to the database."*

We remark that by the triangulation theorem (Theorem 13.26), the existence of such semi-linear set is guaranteed.

So, given a database D, we first apply the linearization query and then apply the `connected` query. This clearly results in the desired $\mathrm{FO}(+, \times, <, 0, 1)$+TC formula expressing connectivity.

So far, we have shown individual queries which can be expressed in $\mathrm{FO}(+, \times, <, 0, 1)$+TC. Moreover, we have seen that not all formulas in $\mathrm{FO}(+, \times, <, 0, 1)$+TC are terminating. We now describe two ways of controlling this termination and its effect on the expressiveness.

Transitive Closure with stop conditions. A formula in $\mathrm{FO}(+, \times, <, 0, 1)$ +TCS is built in the same way as an $\mathrm{FO}(+, \times, <, 0, 1)$ formula, but with the following extra formation rule: if $\psi(\vec{x}, \vec{y})$ is a formula with $\vec{x}$ and $\vec{y}$ k-tuples of real variables, σ is an $\mathrm{FO}(+, \times, <, 0, 1)$ sentence over the input database and a special $2k$-ary relation name X, and $\vec{s}$, $\vec{t}$ are k-tuples of real variables, then

(13.2) $$[\mathrm{TC}_{\vec{x};\vec{y}}\, \psi(\vec{x},\vec{y}) \mid \sigma](\vec{s},\vec{t})$$

is also a formula which has as free variables those in $\vec{s}$ and $\vec{t}$. We call σ the *stop condition* of this formula.

The semantics of a subformula of the above form (13.2) evaluated on databases D is defined in the same manner as in the case without stop condition, but now we stop not only in case an i is found such that $X_i = X_{i+1}$, but also when an i is found such that $(D, X_{i+1}) \models \sigma$, whichever case occurs first. As above, we also consider the restriction $\mathrm{FO}(+,<,0,1)$+TCS.

EXAMPLE 13.53 As an example of an $\mathrm{FO}(+,\times,<,0,1)$+TCS formula over a two-dimensional input database S, we take

$$[\mathrm{TC}_{x;y}\, S(x,y) \mid \exists x \exists y (X(x,y) \wedge y = 1 \wedge 10x \leq 1)](s,t).$$

Here the stop condition is $\sigma \equiv \exists x \exists y (X(x,y) \wedge y = 1 \wedge 10x \leq 1)$. When applied to the graph of the function shown in Fig. 13.23, we see that X_3 satisfies the sentence in the stop condition since for instance $(\frac{1}{16}, 1)$ belongs to it. The evaluation has become terminating (as opposed to the expression without stop condition in Example 13.49). On input the graph of the function shown in Fig. 13.23, this expression still terminates after four iterations (since $X_5 = X_4$, not because the stop condition is satisfied) and returns the same result as in the case without stop condition.

Transitive Closure with start points and parameters. We can also allow parameters in the transitive closure and restrict the computation of the transitive closure to certain paths, after specifying starting points. We denote the resulting logic with $\mathrm{FO}(+,\times,<,0,1)$+KTC. A formula in $\mathrm{FO}(+,\times,<,0,1)$+KTC is built exactly as in $\mathrm{FO}(+,\times,<,0,1)$+TC with the exception that parameters $\vec{u}$ are allowed, i.e.,

(13.3) $$\left[\mathrm{TC}_{\vec{x};\vec{y}} \psi(\vec{x},\vec{y},\vec{u})\right](\vec{s},\vec{t})$$

has as free variables $\vec{u}$, $\vec{s}$ and $\vec{t}$.

The semantics of a subformula of the form (13.3), with $\vec{s} = (s_1, \ldots, s_k)$, evaluated on a database D is defined in the following operational manner: Let I be the set of indices i for which s_i is a constant. Then we start computing the following iterative sequence of $(2k+\ell)$-ary relations:

$$X_0 := \psi(D) \wedge \bigwedge_{i \in I} (s_i = x_i)$$

and

$$X_{i+1} := X_i \cup \{(\vec{x},\vec{y},\vec{u}) \in \mathbb{R}^{2k+\ell} \mid \exists \vec{z}\, (X_i(\vec{x},\vec{z},\vec{u}) \wedge \psi(\vec{z},\vec{y},\vec{u}))\}$$

and stop as soon as $X_i = X_{i+1}$. The semantics of

$$[\mathrm{TC}_{\vec{x};\vec{y}}\, \psi(\vec{x}, \vec{y}, \vec{u})](\vec{s}, \vec{t})$$

is then defined as $(\vec{s}, \vec{t}, \vec{u})$ belonging to the $(2k + \ell)$-ary relation X_i.

EXAMPLE 13.54 As an example of an $\mathrm{FO}(+, \times, <, 0, 1)$+KTC formula over a two-dimensional input database D, we take

$$[\mathrm{TC}_{x;y}\, S(x, y)](\frac{1}{4}, t).$$

When applied to the graph of the function, shown in Fig. 13.23, we see that $X_0 = D \cap \{(x, y) \mid x = \frac{1}{4}\}$ and this set is just $\{(\frac{1}{4}, \frac{1}{2})\}$. Next, X_1 is computed to be $\{(\frac{1}{4}, \frac{1}{2})\} \cup \{(\frac{1}{4}, 1)\}$. In subsequent iterations, no further tuples are added (i.e., $X_2 = X_1$). This example shows that in $\mathrm{FO}(+, \times, <, 0, 1)$+KTC, the evaluation can be restricted to the computation of certain paths in the transitive closure and this gives control over the termination.

We will also consider the fragment $\mathrm{FO}(+, <, 0, 1)$+KTC of this language.

6.2 Computational Completeness

One may wonder what the expressive power of the transitive-closure logics is. It turns out that adding stop conditions or allowing start points and parameters drastically increases the expressive power. More specifically, both $\mathrm{FO}(+, <, 0, 1)$+TCS and $\mathrm{FO}(+, <, 0, 1)$+KTC are computationally complete on semi-linear constraint databases.

This means that for every partial computable (in the sense of Turing computable) query Q on semi-linear databases, there exists a formula φ in $\mathrm{FO}(+, <, 0, 1)$+TCS (respectively in $\mathrm{FO}(+, \times, <, 0, 1)$+KTC) such that for each semi-linear database D, $\varphi(D)$ is defined if and only if $Q(D)$ is defined, and in this case $\varphi(D)$ and $Q(D)$ are equal.

THEOREM 13.55 *Both* $\mathrm{FO}(+, <, 0, 1)$+TCS *and* $\mathrm{FO}(+, <, 0, 1)$+KTC *are computationally complete on linear constraint databases.*

We remark that this holds only for linear databases. However, a similar result holds on arbitrary databases for the extension $\mathrm{FO}(+, \times, <, 0, 1)$+WHILE of $\mathrm{FO}(+, \times, <, 0, 1)$ with a while-loop.

An extension of $\mathrm{FO}(+, \times, <, 0, 1)$ **with a while-loop.** A program in $\mathrm{FO}(+, \times, <, 0, 1)$+WHILE is a finite sequence of *assignment statements* and *while-loops*. Each assignment statement has the form

$$R := \{(x_1, \ldots, x_k) \mid \varphi(x_1, \ldots, x_k)\},$$

where φ is an $\mathrm{FO}(+, \times, <, 0, 1)$ formula that uses the relation names S_i (of the input database schema $\mathcal{S} = \{S_1, \ldots, S_n\}$) and previously introduced relation names. Each while-loop has the form

$$\text{WHILE } \varphi \text{ DO } P \text{ OD},$$

where P is a $\mathrm{FO}(+, \times, <, 0, 1)$+WHILE program and φ an $\mathrm{FO}(+, \times, <, 0, 1)$ formula that uses relation names S_i from the input schema and previously introduced relation names.

The semantics of a program applied to a spatial databases is the operational, step by step execution. So, the effect of an assignment statement is to evaluate the $\mathrm{FO}(+, \times, <, 0, 1)$ formula on the right-hand side on the constraint database D augmented with the previously assigned-to relation variables, and to assign the result of the evaluation to the relation variable on the left-hand side. The effect of a while-loop is to execute the body as long as non-halting condition φ evaluates to true. One relation name R_{out} is designated as the output relation and when the $\mathrm{FO}(+, \times, <, 0, 1)$+WHILE program terminates, the current value of R_{out} is considered to be the output.

Again, we have the problem of non-terminating while loops, as the following example shows.

EXAMPLE 13.56 Let $D{=}[0, 1]$ and consider the following $\mathrm{FO}(+, \times, <, 0, 1)$+WHILE program P over the input schema $\mathcal{S} = \{S\}$, with S unary:

$R := \{(x) \mid S(x)\};$
$Y := \emptyset;$
WHILE $R \neq \emptyset$ DO
    $Y := \{(x) \mid \varphi(x, R)\};$
    $R := R \setminus Y;$
OD

where φ is an $\mathrm{FO}(+, \times, <, 0, 1)$ such that when it is evaluated on any finite sequence of closed non-overlapping intervals $\big\{[a_i, b_i]\big\}$ it results in the sequence of closed non-overlapping intervals $\big\{[a_i, a_i + \frac{b_i - a_i}{3}], [a_i + \frac{2(b_i - a_i)}{3}]]\big\}$. Obviously, P is non-terminating on D and is $P(D)$ is undefined. However, if we allowed infinitely long computations, then $P(D)$ would be the Cantor set which is depicted in Fig. 13.28.

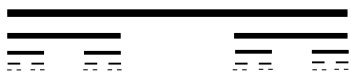


Figure 13.28. The iterative construction of the Cantor set.

The previous example shows that in contrast to the transitive-closure logics, an $\mathrm{FO}(+,\times,<,0,1)$+WHILE program can recursively reduce the size of relations. This already hints that $\mathrm{FO}(+,\times,<,0,1)$+WHILE is more powerful. Indeed, we have the following result.

THEOREM 13.57 *The languages* $\mathrm{FO}(+,\times,<,0,1)$+WHILE *is computationally complete on constraint databases.*

The only result known for transitive-closure logics on arbitrary databases follows from the completeness on linear databases (Theorem 13.55) and the fact the linearization query is expressible in $\mathrm{FO}(+,\times,<,0,1)$+TC (Theorem 13.52):

THEOREM 13.58 *The languages* $\mathrm{FO}(+,\times,<,0,1)$+TCS *and* $\mathrm{FO}(+,\times,<,0,1)$+KTC *are computationally complete on constraint databases as far as Boolean topological queries is concerned.*

Indeed, given a partial computable Boolean topological query Q on a database D, we can apply the linearization query to get $\widehat{D}$, that contains linearizations of all relations in D. By definition of a topological query, $Q(D)$ and $Q(\widehat{D})$ are equal. Since there exists an $\mathrm{FO}(+,\times,<,0,1)$+TCS(respectively $\mathrm{FO}(+,\times,<,0,1)$+KTC) formula φ expressing Q on linear database, we can express Q on D in $\mathrm{FO}(+,\times,<,0,1)$+TCS(respectively $\mathrm{FO}(+,\times,<,0,1)$+KTC) by $\varphi(\widehat{D})$.

There are many open problems related to these recursive extensions of $\mathrm{FO}(+,\times,<,0,1)$. For example, it is not known whether $\mathrm{FO}(+,\times,<,0,1)$+TC is less expressive than $\mathrm{FO}(+,\times,<,0,1)$+TCS or $\mathrm{FO}(+,\times,<,0,1)$+KTC and it is also unknown whether $\mathrm{FO}(+,\times,<,0,1)$+TCS and $\mathrm{FO}(+,\times,<,0,1)$+KTC are complete on arbitrary databases.

So far, we have been extending $\mathrm{FO}(+,\times,<,0,1)$ with generic operators expressing many (in some case every) queries inexpressible in $\mathrm{FO}(+,\times,<,0,1)$. As a result, queries could be undefined because of non-terminating computations.

In the next section we will show how $\mathrm{FO}(+,\times,<,0,1)$ can be extended with specific operators aimed to express specific queries and at the same time guarantee closure.

6.3 Extensions of $\mathrm{FO}(+,\times,<,0,1)$ with topological operators

We have seen that connectivity of a databases is a property which one would like to see being expressible in a query language. We have seen that $\mathrm{FO}(+,\times,<,0,1)$ lacks the power to express this query, while $\mathrm{FO}(+,\times,<,0,1)$+TC provides enough power to do so.

However, it is possible to extend $\mathrm{FO}(+, \times, <, 0, 1)$ directly with a connectivity operator. More generally, define a topological property Top as a collection $\{\mathcal{T}_1, \ldots, \mathcal{T}_n, \ldots\}$ where $\mathcal{T}_n$ is a family of sets in $\mathbb{R}^n$ such that if $X \in \mathcal{T}_n$, then for each homeomorphism
h of $\mathbb{R}^n$, $h(X) \in \mathcal{T}_n$.

EXAMPLE 13.59 Of special interest is when Top expresses the property of being connected. Other examples of Top are the property of being closed, being of dimension k, containing exactly two holes and so on.

Let $\mathbb{T}$ be a set of topological properties. We then define the language $\mathrm{FO}(+, \times, <, 0, 1) + \mathbb{T}$ by extending $\mathrm{FO}(+, \times, <, 0, 1)$ with the following rule: if $\varphi(\vec{x}, \vec{y})$ is a query then

$$\psi(\vec{x}) \equiv \mathsf{Top}\vec{y} \odot \varphi(\vec{x}, \vec{y})$$

is also a query for any $\mathsf{Top} \in \mathbb{T}$. The semantics of such a formula is as follows: $D \models \varphi(\vec{a})$ iff $\varphi(\vec{a}, D) = \{\vec{b} \mid D \models \varphi(\vec{a}, \vec{b})\} \in \mathsf{Top}$.

EXAMPLE 13.60 The query "Is the intersection of regions S and T connected" can be written as $\mathsf{Conn}\vec{x} \odot S(\vec{x}) \wedge T(\vec{x})$. The query $\varphi(x) \equiv \mathsf{Conn}(y, z) \odot S(x, y, z)$ returns the set of all $c \in \mathbb{R}$ for which the intersection of S with the plane $x = c$ is a connected set.

THEOREM 13.61 *The logic* $\mathrm{FO}(+, \times, <, 0, 1) + \mathbb{T}$ *is closed on constraint databases.*

Proof The result follows from a simple induction on the formulas. The only case to prove is $\psi(\vec{x}) \equiv \mathsf{Top}\vec{y} \odot \varphi(\vec{x}, \vec{y})$ for $\mathsf{Top} \in \mathbb{T}$. Let $\vec{x}$ and $\vec{y}$ be of length n and m, respectively. By induction, $\varphi(D)$ results in a semi-algebraic set in $\mathbb{R}^n \times \mathbb{R}^m$. By the triviality theorem for semi-algebraic sets, there exists a decomposition $\mathcal{C}$ of $\mathbb{R}^{n+m}$ into finitely many cells which is trivial over $\mathbb{R}^n$ and such that $\varphi(D)$ is the union of cells of $\mathcal{C}$. Let $\mathcal{C}'$ be the projection of $\mathcal{C}$ onto $\mathbb{R}^n$, and let C be a cell in $\mathcal{C}'$. By triviality, for every $\vec{a}, \vec{b} \in C$, it is the case that $(\varphi(D))_{\vec{a}}$ and $(\varphi(D))_{\vec{b}}$ are homeomorphic, and thus agree on Top. Therefore, the output of ψ on D is a union of finitely many cells in $\mathcal{C}'$. Moreover, since each cell is semi-algebraic, so will be the output $\psi(D)$. QED

The linear analog of the previous theorem also holds for the language $\mathrm{FO}(+, <, 0, 1) + \mathbb{T}$. However, since the triviality theorem does not hold in this case, one first needs to develop an alternative decomposition theorem (see Benedikt et al., 2003 for details).

References

Afrati, F., Cosmadakis, S., Grumbach, S., and Kuper, G. (1994). Linear versus polynomial constraints in database query languages. In Borning, A., editor, *Proceedings of the 2nd Workshop on Principles and Practice of Constraint Programming*, volume 874 of *Lecture Notes in Computer Science*, pages 181–192, Berlin. Springer-Verlag.

Andradas, C., Bröcker, L., and Ruiz, J. M. (1996). *Constructible sets in real geometry*, volume 33 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Basu, S., Pollack, R., and Roy, M.-F. (2003a). *Algorithms in real algebraic geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer-Verlag.

Basu, S., Pollack, R., and Roy, M.-F. (2003b). *Algorithms in Real Algebraic Geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer-Verlag.

Belegradek, O. V., Stolboushkin, A. P., and Taitslin, M. A. (1996). On order-generic queries. Technical Report 96-01, DIMACS.

Benedetti, R. and Risler, J.-J. (1990). *Real algebraic and semi-algebraic sets*. Actualités Mathématiques. [Current Mathematical Topics]. Hermann.

Benedikt, M., Dong, G., Libkin, L., and Wong, L. (1996). Relational expressive power of constraint query languages. In *Proceedings of the 15th ACM Symposium on Principles of Database Systems*, pages 5–16.

Benedikt, M., Grohe, M., Libkin, L., and Segoufin, L. (2003). Reachability and connectivity queries in constraint databases. *J. Comput. System Sci.*, 66(1):169–206.

Benedikt, M. and Keisler, H. J. (2000). Definability over linear constraints. In Clote, P. and Schwichtenberg, H., editors, *Proceedings of Computer Science Logic, 14th Annual Conference of the EACSL*, volume 1862 of *Lecture Notes in Computer Science*, pages 217–231. Springer-Verlag.

Benedikt, M., Kuijpers, B., Löding, C, Van den Bussche, J., and Wilke, T. (2006). A characterization of first-order topological properties of planar spatial data. *Journal of the ACM*.

Benedikt, M. and Libkin, L. (1996). On the structure of queries in constraint query languages. In *11th Annual IEEE Symposium on Logic in Computer Science*, pages 25–34.

Benedikt, M. and Libkin, L. (2000). Safe constraint queries. *SIAM J. Comput.*, 29(5):1652–1682.

Benedikt, M.A. and Libkin, L. (1997). Languages for relational databases over interpreted structures. In *Proceedings of the 16th ACM Symposium on Principles of Database Systems*, pages 87–98.

Bochnak, J., Coste, M., and Roy, M.-F. (1987). *Géométrie algébrique réelle*. Springer-Verlag.

Bochnak, J., Coste, M., and Roy, M.-F. (1998). *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag.

Caviness, B.F. and Johnson, J.R. (1998). *Quantifier Elimination and Cylindrical Algebraic Decomposition*. New York: Springer-Verlag.

Codd, E. (1970). A relational model for large shared databanks. *Communications of the ACM*, 13(6):377–387.

Collins, G.E. (1975). Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In Brakhage, H., editor, *Automata Theory and Formal Languages*, volume 33 of *Lecture Notes in Computer Science*, pages 134–183, Berlin. Springer-Verlag.

Coste, M (2000a). *An Introduction to O-minimal Geometry*. Istituti Editoriali e Poligrafici Internazionali, Pisa.

Coste, M (2000b). *An Introduction to Semialgebraic Geometry*. Istituti Editoriali e Poligrafici Internazionali, Pisa.

Ebbinghaus, H.-D. and Flum, J. (1995). *Finite model theory*. Perspectives in Mathematical Logic. Springer-Verlag.

Ebbinghaus, H.-D., Flum, J., and Thomas, W. (1984). *Mathematical Logic*. Undergraduate Texts in Mathematics. Springer-Verlag.

Geerts, F. (2003). Expressing the box cone radius in the relational calculus with real polynomial constraints. *Discrete Comput. Geom.*, 30(4):607–622.

Geerts, F. and Kuijpers, B. (2000). Linear approximation of planar spatial databases using transitive-closure logic. In *Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 126–135.

Geerts, F. and Kuijpers, B. (2005). On the decidability of termination of query evaluation in transitive-closure logics for polynomial constraint databases. *Theoretical Computer Science*, 336(1):125–151. Database Theory–Special issue with selected papers of ICDT'03.

Giusti, M., Lecerf, G., and Salvy, B. (2001). A Gröbner free alternative for polynomial system solving. *Journal of Complexity*, 17(1):154–211.

Grohe, M. and Segoufin, L. (2002). On first-order topological queries. *ACM Transactions on Computational Logic*, 3(3):336–358.

Grumbach, S. and Su, J. (1995). First-order definability over constraint databases. In *Proceedings of 1st Conference on Principles and Practice of Constraint Programming*, volume 976 of *Lecture Notes in Computer Science*. Springer-Verlag.

Grumbach, S., Su, J., and Tollu, C. (1995). Linear constraint query languages: expressive power and complexity. In Leivant, D., editor, *Logic and*

Computational Complexity, volume 960 of *Lecture Notes in Computer Science*, pages 426–446. Springer-Verlag.

Gyssens, M., Van den Bussche, J., and Van Gucht, D. (1999). Complete geometric query languages. *J. Comput. System Sci.*, 58(3):483–511.

Heintz, J., Roy, M.-F., and Solernó, P. (1993). Description of the connected components of a semialgebraic set in single exponential time. *Discrete and Computational Geometry*, 6:1–20.

Hong, H. (1990). QEPCAD — quantifier elimination by partial cylindrical algebraic decomposition. http://www.cs.usna.edu/~qepcad/B/QEPCAD.html.

Kanellakis, P. C., Kuper, G., and Revesz, P. Z. (1995). Constraint query languages. *Journal of Computer and System Sciences*, 51:26–52.

Kreutzer, S. (2001). Operational semantics for fixed-point logics on constraint databases. In *Logic for programming, artificial intelligence, and reasoning*, volume 2250 of *Lecture Notes in Compututer Science*, pages 470–484. Springer-Verlag.

Kuijpers, B., Paredaens, J., and Van den Bussche, J. (2000). Topological elementary equivalence of closed semi-algebraic sets in the real plane. *J. Symbolic Logic*, 65(4):1530–1555.

Kuper, G. M., Libkin, L., and Paredaens, J., editors (2000). *Constraint Databases*. Springer-Verlag.

Motzkin, T. S. (1936). *Beiträge zur Theorie der linearen Ungleichungen*. Doctoral dissertation. Universität Zürich.

Paredaens, J., Van den Bussche, J., and Van Gucht, D. (1994). Towards a theory of spatial database queries. In *Proceedings of the Thirteenth ACM Symposium on Principles of Database Systems*, pages 279–288.

Paredaens, J., Van den Bussche, J., and Van Gucht, D. (1995). First-order queries on finite structures over the reals. In *Proceedings of the 10th IEEE Symposium on Logic in Computer Science*, pages 79–89.

Revesz, R. Z. (2002). *Introduction to Constraint Databases*. Springer-Verlag.

Rigaux, Ph., Scholl, M., and Voisard, A. (2000). *Introduction to Spatial Databases: Applications to GIS*. Morgan Kaufmann.

Seidenberg, A. (1954). A new decision method for elementary algebra. *Ann. of Math. (2)*, 60:365–374.

Stolboushkin, A.P. and Taitslin, M.A. (1996). Linear vs. order constraints over rational databases. In *Proceedings of the 15th ACM Symposium on Principles of Database Systems*, pages 17–27.

Tarski, A. (1948). *A Decision Method for Elementary Algebra and Geometry*. University of California Press.

TERA-project (1993). http://tera.medicis.polytechnique.fr/index.html.

van den Dries, L. (1998). *Tame Topology and O-minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series*. Cambridge University Press.

Vandeurzen, L., Gyssens, M., and Van Gucht, D. (1996). On query languages for linear queries definable with polynomial constraints. In Freuder, E. F., editor, *Proceedings of the 2nd Conference on Principles and practice of constraint programming*, volume 1118 of *Lecture Notes in Computer Science*, pages 468–481, Berlin. Springer-Verlag.

Chapter 14

MATHEMATICAL MORPHOLOGY

Isabelle Bloch
École Nationale Supérieure des Télécommunications

Henk Heijmans
Centrum voor Wiskunde en Informatica, Amsterdam

Christian Ronse
CNRS-Université Louis Pasteur, Strasbourg I

Second Reader

Johan van Benthem
University of Amsterdam & Stanford University

1. Introduction

Mathematical morphology (MM) is a branch of image processing, which arose in 1964. It is associated with the names of Georges Matheron and Jean Serra, who developed its main concepts and tools, expounded it in several books (Matheron, 1975; Serra, 1982; Serra, 1988), and created a team at the *Centre de Morphologie Mathématique* on the Fontainebleau site of the Paris School of Mines.

MM truly deserves the adjective "mathematical", as it is heavily mathematized. In this respect, it contrasts with the various heuristic or experimental approaches to image processing that one sees in the literature. It stands also as an alternative to another strongly mathematized branch of image processing, the one that bases itself on signal processing and information theory, following the works of prestigious pioneers named Wiener, Shannon, Gabor, etc. Indeed, these classical approaches proved their value in telecommunications. However

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 857–944.
© 2007 *Springer.*

MM claims that analysing the information of an image is not like transmitting a signal on a channel, that an image should not be considered as a combination of sinusoidal frequencies, nor as the result of a Markov process on individual points. It considers that the purpose of image analysis is to find spatial objects, therefore images contain geometrical shapes with luminance (or colour) profiles, which can be investigated by their interactions with other shapes and luminance profiles. This makes the morphological approach especially relevant in situations where image grey-levels (or colours) correspond directly to significant material data, as in medical imaging, microscopy, industrial inspection and remote sensing.

In its development, MM has borrowed concepts and tools from various branches of mathematics: algebra (lattice theory), topology, discrete geometry, integral geometry, geometrical probability, partial differential equations, etc.; in fact any mathematical theory that deals with shapes, their combinations or their evolution, can be brought to contribute to morphological theory.

MM started by analysing binary images (sets of points) with the use of set-theoretical operations. In order to apply it to other types of images, for example grey-level ones (numerical functions), it was necessary to generalize set-theoretical notions, such as the relation of inclusion and the operations of union and intersection. This was done by using the lattice-theoretical notions of a *partial order* relation between images, for which the operations of *supremum* (least upper bound) and *infimum* (greatest lower bound) are defined. Therefore the central structure in MM is that of a *complete lattice*, and the basic morphological operators (dilation, erosion, opening and closing) can be characterized in this framework.

When analysing sets, one considers their topology: is the set in one or several pieces, how many holes has it, etc. Some topological notions, in particular connectedness, have been generalized in the framework of complete lattices. Nowadays, most morphological techniques combine lattice-theoretical and topological methods.

The computer processing of pictures quickly led to digital models of geometry. The pioneering work in this field is that of Azriel Rosenfeld, who died in 2004 after having contributed to digital geometry and image processing for 40 years. Thanks to its algebraic formalism, mathematical morphology is perfectly adapted to the digital framework. Moreover, the topology of digital figures can be studied in the framework of *combinatorial topology*, a field that was developed in the first half of the 20th century by mathematicians like Paul Alexandroff (Alexandroff, 1937; Alexandroff, 1956; Alexandroff and Hopf, 1935). In particular the latter proposed in 1935 to subdivide the Euclidean plane into rectangular cells, in such a way that cell interiors, sides, and corners are considered as points in an abstract space, whose combinatorial relations provide the topology. This idea prefigured the notion of pixels, and the corresponding

Alexandroff topology was formally developed by Efim Khalimsky and popularized by Vladimir Kovalevsky; it has been shown that many "paradoxes" of digital geometry (like non-parallel lines which do not intersect) find a natural solution in that topology.

MM has also borrowed tools from integral geometry in order to measure some parameters on images. However these measurements are usually preceded by some image processing operations, in order to restrict the measure to some appropriate features: for example, to estimate the average length of particles whose width is at least w, we apply first an operator eliminating all particles narrower than w, then we make a length measurement on the remaining ones.

MM has also a probabilistic aspect, where images and shapes can be considered as random events. Suppose for example that one asks n experts to extract a certain set S from an image, say n anatomists have to extract the left half of the liver from an X-ray scanner hepatic image; they will disagree, and extract n different sets $S_1, \ldots, S_n$; now, how does one derive the "average" of these n sets, or their "standard deviation"? Furthermore, if one designs a computer algorithm for extracting that set, which produces the set S_{auto}, how does on evaluate the statistical significance of S_{auto} w.r.t. to the distribution $S_1, \ldots, S_n$? Such problems are studied in geometric probability, through the theory of random sets and functions (Matheron, 1975; Serra, 1982; Serra, 1988). This should not be confused with Markov field models for image processing: there the random variable is the grey-level of an individual pixel, and it evolves in space by a Markov process.

Image analysis has considered the varying scales at which things are seen. This has been formalized by multi-scale models governed by partial differential equations (PDEs). This has happened also for morphological operators, for which new PDEs have been given, leading to a new understanding of their functioning.

The theory of morphological operators relies on the formalism of lattice theory, and the latter underlies also several theoretical aspects of computer science: fuzzy sets, formal concept analysis and abstract interpretation of programming. In fact, the lattice-theoretical tools developed in each speciality can be used for the other ones. For example, a research on fuzzy morphology has been undertaken since several years. Also, the tools of MM, developed for the purpose of filtering and segmenting images, have found applications for modelling spatial concepts, like "close to" or "between".

The link between logic and lattice theory is obvious. Boole's logic is the first example of a Boolean algebra, while non-classical logics have been modeled as non-Boolean lattices. As MM analyses spatial shapes by means of lattice-theoretical operations, it is adapted to the logical analysis of spatial relations, while its abstract mathematical tools can be used in order to illuminate some

aspects of logic, for example modal logic, and to build new operations in such a framework.

The purpose of this chapter is to present the basic theory of MM (Sec. 2), then to show how its tools can be applied to various specialities dealing with the analysis of spatial shapes and spatial relations, such as formal concept analysis, rough sets and fuzzy sets (Sec. 3), and finally to show its relevance in logic (Sec. 4).

Let us now describe the basic operations of mathematical morphology, first in the case of sets (or binary images), and next in the case of numerical functions (or grey-level images). We must warn the reader that in several works (including important ones, for instance Serra, 1982; Soille, 2003), the definitions given for the basic operations (Minkowski addition and subtraction, dilation, erosion, opening and closing) differ from ours in that in some cases the structuring element must be replaced by its symmetrical; also the notation can be different (in particular Serra, 1982). The definitions given here for morphological operations are standard (Heijmans, 1994), in the sense that they are consistent with the original definitions given by Minkowski, 1903 for the Minkowski addition and Hadwiger, 1950 for the Minkowski subtraction, and that they follow the algebraic theory (see Sec. 2), which allows to give a unified treatment (Heijmans and Ronse, 1990; Ronse and Heijmans, 1991) of such operators in the case of sets, numerical functions, and many other structures.

1.1 Morphology on sets

Consider the space $E = \mathbb{R}^n$ or $\mathbb{Z}^n$, with origin $o = (0, \ldots, 0)$. Given $X \subseteq E$, the *complement* of $X \subseteq E$ is $X^c = E \setminus X$, and the *transpose* or *symmetrical* of X is $\check{X} = \{-x \mid x \in X\}$. For every $p \in E$, the *translation* by p is the map $E \to E : x \mapsto x + p$; it transforms any subset X of E into its *translate by* p, $X_p = \{x + p \mid x \in X\}$.

Most morphological operations on sets can be obtained by combining set-theoretical operations with two basic operators, *dilation* and *erosion*. The latter arise from two set-theoretical operations, the *Minkowski addition* $\oplus$ (Minkowski, 1903) and *subtraction* $\ominus$ (Hadwiger, 1950), defined as follows for any $X, B \in \mathcal{P}(E)$:

$$
\begin{aligned}
X \oplus B &= \bigcup_{b \in B} X_b \ , \\
&= \bigcup_{x \in X} B_x \ , \\
&= \{x + b \mid x \in X, b \in B\} \ ; \\
X \ominus B &= \bigcap_{b \in B} X_{-b} \ , \\
&= \{p \in E \mid B_p \subseteq X\} \ .
\end{aligned}
\tag{14.1}
$$

Formally speaking, X and B play similar roles as binary operands. However, in real situations, X will stand for the *image* (which is big, and given by the problem), and B for the *structuring element* (a small shape chosen by the user), so that $X \oplus B$ and $X \ominus B$ will be transformed images. We define the *dilation by* B, $\delta_B : \mathcal{P}(E) \to \mathcal{P}(E) : X \mapsto X \oplus B$, and the *erosion by* B, $\varepsilon_B : \mathcal{P}(E) \to \mathcal{P}(E) : X \mapsto X \ominus B$. It should be noted that dilation and erosion are *dual by complementation*, in other words dilating a set is equivalent to eroding its complement with the symmetrical structuring element:

$$(X \oplus B)^c = X^c \ominus \check{B} \ , \qquad (X \ominus B)^c = X^c \oplus \check{B} \ . \tag{14.2}$$

Therefore the properties of erosion are derived from those of dilation by duality: dilation inflates the object, deflates the background and deforms convex corners of the object; thus erosion deflates the object, inflates the background and deforms concave corners of the object. By Equation (14.2), we can also obtain alternate formulations for Minkowski addition and subtraction:

$$\begin{array}{rcl} X \oplus B & = & \{p \in E \mid (\check{B})_p \cap X \neq \emptyset\} \ ; \\ X \ominus B & = & \{p \in E \mid \forall z \notin X, p \notin (\check{B})_z\} \ . \end{array} \tag{14.3}$$

We illustrate in Fig. 14.1 the dilation and erosion of a cross by a triangular structuring element.

Dilation and erosion are the basic elements from which most morphological operators are built. The first example is the *hit-or-miss transform*, which uses a pair of structuring elements. Let A and B be two disjoint subsets of E; A will be the *foreground structuring element* and B the *background structuring element*; we then define:

$$\begin{array}{rcl} X \circledast (A, B) & = & \{p \in E \mid A_p \subseteq X \text{ and } B_p \subseteq X^c\} \ , \\ & = & (X \ominus A) \cap (X^c \ominus B) = (X \ominus A) \setminus (X \oplus \check{B}) \ . \end{array}$$

This will give the locus of all points where A fits the foreground and B fits the background. This operation corresponds to what is usually called *template matching*.

The main operators derived from dilation and erosion are *opening* and *closing*. We define the binary operations $\circ$ and $\bullet$ by setting for any $X, B \in \mathcal{P}(E)$:

$$\begin{array}{rcl} X \circ B & = & (X \ominus B) \oplus B \ , \\ & = & \bigcup \{B_p \mid p \in E \text{ and } B_p \subseteq X\} \ ; \\ X \bullet B & = & (X \oplus B) \ominus B \ . \end{array} \tag{14.4}$$

The operator $\gamma_B : \mathcal{P}(E) \to \mathcal{P}(E) : X \mapsto X \circ B$ is called the *opening by* B; it is the composition of the erosion ε_B, followed by the dilation δ_B. On the other hand, the operator $\varphi_B : \mathcal{P}(E) \to \mathcal{P}(E) : X \mapsto X \bullet B$ is called the *closing*

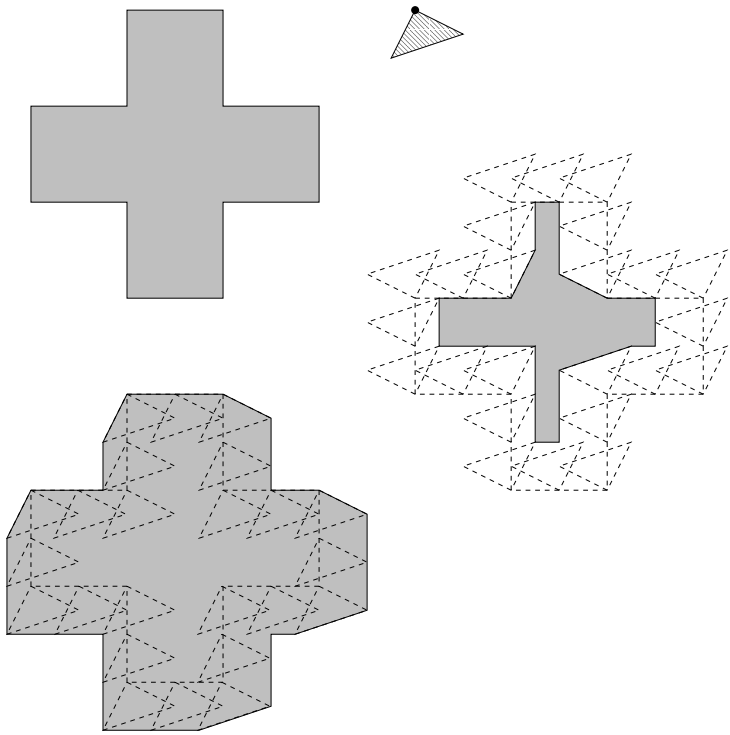


Figure 14.1. Top: The figure X is the cross, and the structuring element B is the triangle; the position of the origin is indicated by a thick dot. Bottom: The dilation $X \oplus B$ of X by B. Right: The erosion $X \ominus B$ of X by B is obtained as the complement of the dilation $X^c \oplus \check{B}$ of the complement X^c by the symmetrical structuring element $\check{B}$.

by B; it is the composition of the dilation δ_B, followed by the erosion ε_B. The two are dual by complementation:

$$(X \circ B)^c = X^c \bullet \check{B} \ , \qquad (X \bullet B)^c = X^c \circ \check{B} \ . \tag{14.5}$$

Hence the properties of closing are derived from those of opening by duality: opening removes narrow parts of the object and deforms convex corners of the object; thus closing fills narrow parts of the background and deforms concave corners of the object. We illustrate in Fig. 14.2 the opening and closing of a cross by a triangular structuring element.

Given a family $\mathcal{B}$ of structuring elements, the opening by $\mathcal{B}$, written $\gamma_{\mathcal{B}}$, is the union of openings by elements of $\mathcal{B}$, while the closing by $\mathcal{B}$, written $\varphi_{\mathcal{B}}$, is

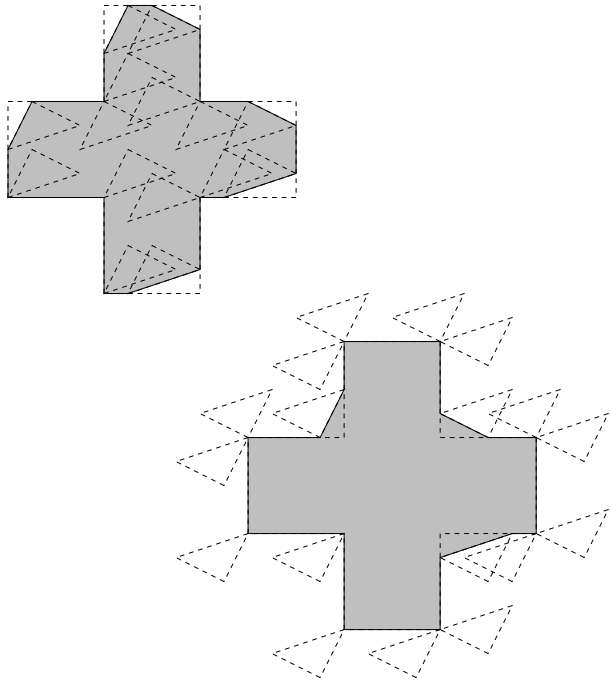


Figure 14.2. We use the same figure X and structuring element B as in Fig. 14.1. Top left: The opening $X \circ B$ of X by B is the union of all translates of B which are included in X. Bottom right: The closing $X \bullet B$ of X by B is obtained as the complement of the opening $X^c \circ \check{B}$ of the complement X^c by the symmetrical structuring element $\check{B}$.

the intersection of closings by elements of $\mathcal{B}$:

$$\begin{aligned} \gamma_{\mathcal{B}}(X) &= \bigcup_{B \in \mathcal{B}} (X \circ B) \ , \\ \varphi_{\mathcal{B}}(X) &= \bigcap_{B \in \mathcal{B}} (X \bullet B) \ . \end{aligned} \tag{14.6}$$

For example, if H and V are respectively a horizontal and a vertical line segment of length a, $\gamma_{\{H,V\}}$ will extract from a line drawing all horizontal and vertical lines of length at least a (as well as all blobs whose height or width is at least a).

The most interesting properties of the opening and closing (by one or several structuring elements) is that they are *idempotent*: $\gamma_{\mathcal{B}}(\gamma_{\mathcal{B}}(X)) = \gamma_{\mathcal{B}}(X)$ and $\varphi_{\mathcal{B}}(\varphi_{\mathcal{B}}(X)) = \varphi_{\mathcal{B}}(X)$. This means that if we consider them as filters, they do their job completely, and there is no need to repeat them. This contrasts with the behaviour of other image processing operators, like the median filter, where repeated applications can further modify the image, without a guarantee that it

will reach a stable result after a finite number of iterations (indeed, the median filter can produce oscillations). The opening can be used to filter out *positive noise*, that is, to remove noisy parts of the object, typically small components; on the other hand, the closing can be used to remove *negative noise*, that is, to add to the object noisy parts of the background, typically small holes. By repeated composition of an opening and a closing, one can obtain four new filters:

- opening followed by closing;
- closing followed by opening;
- opening followed by closing, then by opening;
- closing followed by opening, then by closing.

All four are idempotent, and no other operator can be obtained by further composition (Serra, 1988). They can be used as filters to remove both positive and negative noise; for example, they constitute an alternative to median filtering for removing speckle noise.

These operators have a drawback: they deform the frontier between the object and background. Typically, if one uses a disk-shaped structuring element, they will round the corners of objects. However, one may want to filter out small components or holes of the object, without modifying the shape of the other components and holes. In other words, we look for filters which do not act at the level of pixels, but of connected components of the foreground (called *grains*) and of the background (called *pores*).

The basic operation for this purpose is shown in Fig. 14.3: from a figure F, we extract the union of all connected components of F (grains) that intersect a marker R.

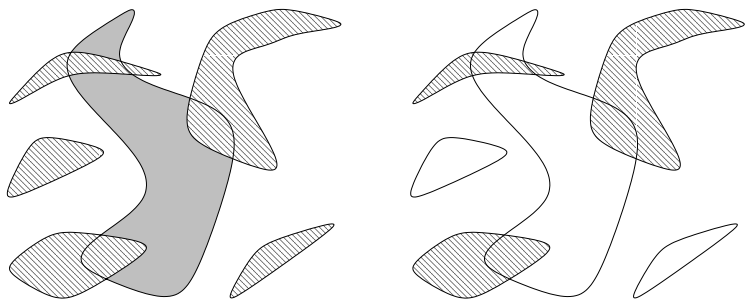


Figure 14.3. Left: We have a figure F (shown hatched) and a marker R (grey). Right: All connected components of F that intersect R are shown hatched.

We can formalize this operation as follows. We assume that E is a digital space ($E = \mathbb{Z}^n$ or a bounded grid in $\mathbb{Z}^n$), and that the connectivity arises from

an adjacency graph on E, for example, the 4- or 8-adjacency on $\mathbb{Z}^2$, the 6- or 26-adjacency on $\mathbb{Z}^3$ (Rosenfeld and Kak, 1976). Let V be the structuring element comprising the origin o and the pixels adjacent to it, so that for any pixel p, the set comprising p and its neighbours is V_p; note that V is symmetrical ($V = \check{V}$). Given a set F (called the *mask*) and a subset R of F (called the *marker*), we define the *geodesical reconstruction by dilation* (from marker R in the mask F) as the limit

$$rec_{\oplus}(F, R) = \bigcup_{n \in \mathbb{N}} R_n$$

of the increasing sequence of sets R_n, $n \in \mathbb{N}$, defined recursively as follows:

$$R_0 = R \cap F \qquad \text{and} \qquad \forall n \in \mathbb{N}, \quad R_{n+1} = (R_n \oplus V) \cap F \ .$$

This will indeed give the union of all grains of F *marked by* (i.e., intersecting) the marker R.

The dual operation is the *geodesical reconstruction by erosion*; here the marker R is a superset of the mask F ($F \subseteq R$), and it is defined as

$$rec_{\ominus}(F, R) = \left[rec_{\oplus}(F^c, R^c)\right]^c \ .$$

This is in fact the limit $\bigcap_{n \in \mathbb{N}} R_n$ of the decreasing sequence of sets R_n, $n \in \mathbb{N}$, defined recursively by

$$R_0 = R \cup F \qquad \text{and} \qquad \forall n \in \mathbb{N}, \quad R_{n+1} = (R_n \ominus V) \cup F \ .$$

The behaviour of $rec_{\ominus}$ is to reconstruct all pores of F which are not completely covered by the marker R; in other words, all connected components of the background F^c which are included in R, are added to F. We illustrate this operation in Fig. 14.4.

Given an opening γ, we define the *opening by reconstruction* γ_{rec} as the geodesical reconstruction by dilation using the opening as marker:

$$\gamma_{rec}(X) = rec_{\oplus}(X, \gamma(X)) \ .$$

Similarly for a closing φ, we define the *closing by reconstruction* φ_{rec} as the geodesical reconstruction by erosion using the closing as marker:

$$\varphi_{rec}(X) = rec_{\ominus}(X, \varphi(X)) \ .$$

Note that for a connected structuring element B containing the origin, we have

$$\begin{aligned} & rec_{\oplus}(X, X \circ B) &= & \ rec_{\oplus}(X, X \ominus B) \\ \text{and} \quad & rec_{\ominus}(X, X \bullet B) &= & \ rec_{\ominus}(X, X \oplus B) \ . \end{aligned}$$

The opening and closing by reconstruction are again idempotent operators; they respectively remove small grains and fill small pores, but they do not deform the

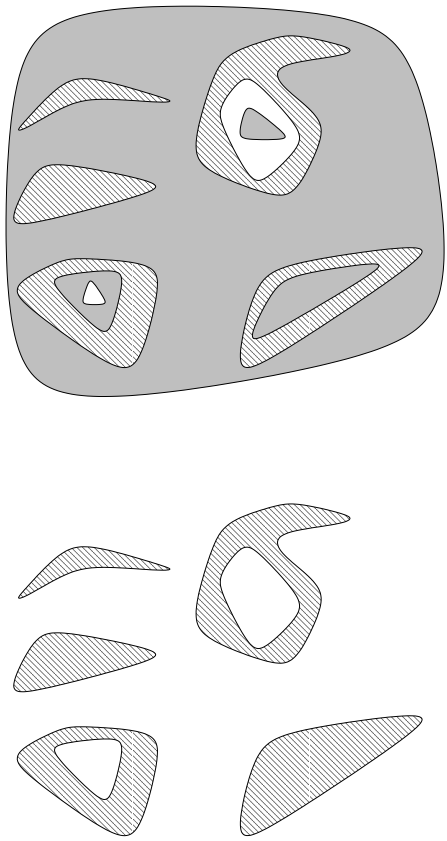


Figure 14.4. Top: We have a mask figure F (hatched) contained in a marker R ($R \setminus F$ is shown in grey). Bottom: $rec_{\ominus}(F, R)$ (hatched) is made of F and all connected components of F^c which are completely covered by the marker R.

remaining boundaries between foreground and background. They can then be composed (as explained above: opening followed by closing, closing followed by opening, etc.) in order to provide idempotent filters that remove grains and pores on the basis of their width, without distorting the contours of objects.

Other idempotent filters can be built, that act directly on grains and pores, for example, the *area opening* (which removes all grains whose area is below a threshold) and the *area closing* (filling all pores whose area is below a threshold).

1.2 Morphology on functions

In computer imaging, grey-levels are coded by numerical values, the low ones corresponding to dark pixels, and the high ones corresponding to bright ones. Hence in mathematical morphology (Heijmans, 1994), grey-level images

are usually considered as numerical functions $E \to T$, where E is the space of points and T is the set of grey-levels; it is always a subset of $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. The grey-levels are numerically ordered, and morphological operations usually compute at each point in E a combination of numerical suprema and infima of grey-level values. Thus one supposes that T is closed under the operations of non-empty numerical supremum and infimum; in the terminology that we will introduce in Sec. 2, T is a *complete lattice*. Usually one takes for T one of the sets $\overline{\mathbb{R}}$, $\overline{\mathbb{Z}} = \mathbb{Z} \cup \{-\infty, +\infty\}$, $[a, b] = \{x \in \overline{\mathbb{R}} \mid a \leq x \leq b\}$ (with $a, b \in \overline{\mathbb{R}}$ and $a < b$), or $[a \ldots b] = [a, b] \cap \overline{\mathbb{Z}}$ (with $a, b \in \overline{\mathbb{Z}}$ and $a < b$). We write t_0 and t_1 respectively for the least and greatest element of T (thus $t_0 = -\infty$ and $t_1 = +\infty$ for $T = \overline{\mathbb{R}}$ or $\overline{\mathbb{Z}}$, while $t_0 = a$ and $t_1 = b$ for $T = [a, b]$ or $[a \ldots b]$).

The set T^E of functions $E \to T$ inherits the numerical order on T by the pointwise ordering of functions:

$$F \leq G \qquad \Longleftrightarrow \qquad \forall p \in E, \quad F(p) \leq G(p) \ . \tag{14.7}$$

This is the analogue for functions of the inclusion relation for sets. Now the analogues for functions of the union and intersection operations for sets, are the *supremum* (least upper bound) and *infimum* (greatest lower bound), obtained by pointwise supremum and infimum operations:

$$\bigvee_{i \in I} F_i : E \to T : p \mapsto \sup_{i \in I} F_i(p) \ , \quad \bigwedge_{i \in I} F_i : E \to T : p \mapsto \inf_{i \in I} F_i(p) \ . \tag{14.8}$$

We write $F \vee G$ and $F \wedge G$ for the supremum and infimum of two functions (cf. the union and intersection of two sets); as the two binary operations $\vee$ and $\wedge$ are commutative and associative, we can write $F_1 \vee \cdots \vee F_n$ and $F_1 \wedge \cdots \wedge F_n$, which are in fact respectively equal to $\bigvee_{i \in \{1,\ldots,n\}} F_i$ and $\bigwedge_{i \in \{1,\ldots,n\}} F_i$. The least and greatest functions are the ones with constant values t_0 and t_1 respectively, they are the analogues of the empty set $\emptyset$ and the whole space E.

Given a function $F : E \to T$ and a point $p \in E$, the *translate of F by p* is the function F_p whose graph is obtained by translating the graph $\{(x, F(x)) \mid x \in E\}$ by p in the first coordinate, that is,

$$\{(y, F_p(y)) \mid p \in E\} = \{(x + p, F(x)) \mid x \in E\} \ ,$$

in other words

$$\forall y \in E, \qquad F_p(y) = F(y - p) \ .$$

We have thus the analogues for functions of the union, intersection and translation operations for sets. It is then possible to define the dilation, erosion, opening and closing of a function by a structuring element, by making analogues of Eqs. (14.1, 14.4).

There is however a systematic method for extending operators on sets to operators on functions (Heijmans, 1991; Heijmans, 1994; Ronse, 2003). It relies on the notions of *thresholding* and *stacking*. Given a function $F : E \to T$, the *umbra* (or *hypograph*) of F is the set

$$U(F) = \{(p,t) \mid p \in E, t \in T, F(p) \geq t\} \ ,$$

and for any value $t \in T$, consider the *threshold set*

$$X_t(F) = \{p \in E \mid F(p) \geq t\} \ ;$$

thus $(p,t) \in U(F)$ iff $p \in X_t(F)$. We illustrate these notions in Fig. 14.5.

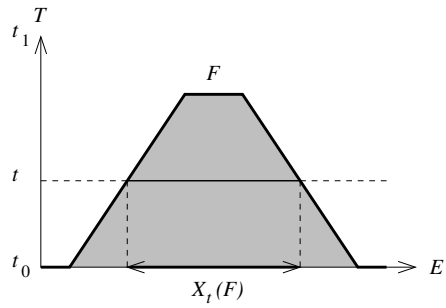


Figure 14.5. The graph of F, and below it the umbra $U(F)$ (in grey). For $t \in T$, the horizontal line at level t crosses the umbra in a section whose projection in E is the threshold set $X_t(F)$.

Given an operator $\psi : \mathcal{P}(E) \to \mathcal{P}(E)$, the *flat operator corresponding to* ψ (or *flat extension of* ψ) is the operator $\psi^T : T^E \to T^E$ constructed as follows:

1 *Thresholding:* For every $t \in T$, we take the horizontal cross-section of the umbra $U(F)$ at level t, that is the set $X_t(F) \times \{t\}$.

2 *Horizontal operation:* We apply ψ horizontally to every such cross-section, that is, for every $t \in T$ we obtain the set $\psi\left(X_t(F)\right) \times \{t\}$.

3 *Stacking:* The upper envelope of these sets $\psi\left(X_t(F)\right) \times \{t\}$, $t \in T$, defines a function which gives $\psi^T(F)$.

We illustrate this construction in Fig. 14.6, in the case where $\psi = \delta_B$, the dilation by a structuring element B. In fact, the values taken by $\psi^T(F)$ are given by the following formula:

$$\forall p \in E, \qquad \psi^T(F)(p) = \bigvee \{t \in T \mid p \in \psi\left(X_t(F)\right)\} \ . \tag{14.9}$$

Rather than using Equation (14.9) to compute the values $\psi^T(F)(p)$, we can rely on the fact that the flat extension of operators transforms the operations on

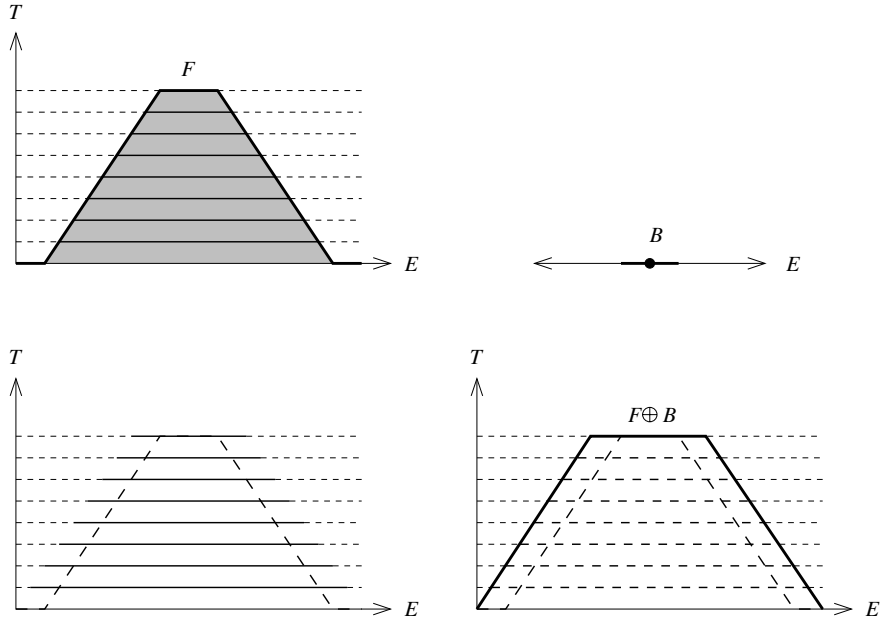


Figure 14.6. Top left: The graph of F, the umbra $U(F)$ (in grey), and horizontal cross-sections $X_t(F) \times \{t\}$ of the umbra. Top right: The structuring element B (the position of the origin is indicated by a dot). Bottom left: We apply δ_B, the dilation by B, horizontally to the cross-sections $X_t(F) \times \{t\}$, obtaining the sets $(X_t(F) \oplus B) \times \{t\}$. Bottom right: The upper envelope of the dilated cross-sections gives the dilated function $\delta_B^T(F)$, also written $F \oplus B$.

sets into the corresponding ones on functions, as it follows from the properties listed below (for the sake of brevity, in the formulas we omit the quantifications $\forall X \in \mathcal{P}(E)$ and $\forall F \in T^E$):

- *Identity*: If $\psi(X) = X$, then $\psi^T(F) = F$.
- *Translation*: If $\psi(X) = X_p$, then $\psi^T(F) = F_p$.
- *Union*: If $\psi(X) = \bigcup_{i \in I} \xi_i(X)$, then $\psi^T(F) = \bigvee_{i \in I} \xi_i^T(F)$.
- *Intersection*: If $\psi(X) = \bigcap_{i \in I} \xi_i(X)$, then $\psi^T(F) = \bigwedge_{i \in I} \xi_i^T(F)$.
- *Composition*: If $\psi(X) = \eta(\zeta(X))$, then $\psi^T(F) = \eta^T(\zeta^T(F))$.

These properties can for example be used to give formulas for the flat extensions of dilation and erosion. As $\delta_B(X) = \bigcup_{b \in B} X_b$ and $\varepsilon_B(X) = \bigcap_{b \in B} X_{-b}$ (see Equation (14.1)), we obtain for every $F \in T^E$:

$$\delta_B^T(F) = \bigvee_{b \in B} F_b \qquad \text{and} \qquad \varepsilon_B^T(F) = \bigwedge_{b \in B} F_{-b} \ . \tag{14.10}$$

We get then for every $p \in E$:

$$\begin{array}{lrcccl} & \delta_B^T(F)(p) & = & \sup_{b\in B} F(p-b) & = & \sup_{q\in(\check{B})_p} F(q) \\ \text{and} & \varepsilon_B^T(F)(p) & = & \inf_{b\in B} F(p+b) & = & \inf_{q\in B_p} F(q) \ . \end{array} \tag{14.11}$$

It is customary to write $F \oplus B$ and $F \ominus B$ for $\delta_B^T(F)$ and $\varepsilon_B^T(F)$. Following Equation (14.4), we define $F \circ B = (F \ominus B) \oplus B$ and $F \bullet B = (F \oplus B) \ominus B$; clearly $F \circ B = \gamma_B^T(F)$ and $F \bullet B = \varphi_B^T(F)$. Note that here the operations $\oplus$, $\ominus$, $\circ$ and $\bullet$ have a function as first operand, a set as second, and a function again as result.

All set operators built by combining dilations and erosions through unions, intersections and translations, extend thus naturally as *flat operators*. Then the properties of the set operators translate directly to their flat extensions; for example, openings and closings are idempotent, and composing them leads to idempotent filters. In practice, flat operators behave on bright and dark parts of a grey-level image in the same way as the corresponding set operators do on foreground and background. For example, dilation inflates bright areas and deflates dark ones, while erosion does the contrary; opening darkens narrow bright zones, while closing brightens narrow dark zones; dilation and opening deform corners which are convex on the bright side, while erosion and closing deform corners which are convex on the dark side. In particular, filters obtained by composing opening and closing can be used to remove small defects in an image, such as speckle noise.

There is still a duality between erosion and dilation, and between opening and closing. Let n be an *inversion* of T, that is a bijection $T \to T$ which reverses the order: $t < t' \iff n(t) > n(t')$; for example, if $T = [a \dots b]$, we have $n(t) = a + b - t$; we extend it to an inversion N on functions, by setting $N(F) : p \mapsto n(F(p))$ (here n and N stand for *negative*, in the photographic sense). Then:

$$\begin{array}{rclrcl} N\big(F \oplus B\big) & = & N(F) \ominus \check{B} \ , & N\big(F \ominus B\big) & = & N(F) \oplus \check{B} \ , \\ N\big(F \circ B\big) & = & N(F) \bullet \check{B} \ , & N\big(F \bullet B\big) & = & N(F) \circ \check{B} \ . \end{array}$$

This expresses formally the fact that the behaviour of erosions and closings is derived of that of dilations and openings, by exchanging the roles of bright and dark points or zones in the grey-level image.

It is also possible to give flat extensions of geodesical reconstruction by dilation or erosion. For a *mask function* and a *marker function* R, such that $R \leq F$, we define the *geodesical reconstruction by dilation*

$$rec_{\oplus}(F, R) = \bigvee_{n\in\mathbb{N}} R_n \ ,$$

where the functions R_n, $n \in \mathbb{N}$, are defined recursively by

$$R_0 = R \wedge F \qquad \text{and} \qquad \forall n \in \mathbb{N}, \quad R_{n+1} = (R_n \oplus V) \wedge F.$$

(Here V is the neighbourhood of the origin.) For $R \geq F$, we have the *geodesical reconstruction by erosion*

$$rec_{\ominus}(F, R) = \bigwedge_{n \in \mathbb{N}} R_n \ ,$$

where

$$R_0 = R \vee F \qquad \text{and} \qquad \forall n \in \mathbb{N}, \quad R_{n+1} = (R_n \ominus V) \vee F \ .$$

In fact, the two are dual:

$$rec_{\ominus}(F, R) = N\big[rec_{\oplus}\big(N(F), N(R)\big)\big] \ .$$

In the same way as the geodesical reconstructions on sets acted on grains and pores (connected components of the foreground and background), here these operators will act on *flat zones*, that is, maximal connected sets having a constant grey-level value. In particular, we can design openings and closings by reconstruction, as in the case of sets, and these filters will remove some bright or dark objects, and simplify the grey-levels of remaining objects, but they will not deform the contours between objects. They are thus very interesting image filters.

The extension of morphology on sets that we have described, is called *flat morphology*. This terminology arises from the fact that we work on the "horizontal" structure of functions (see Fig. 14.6). We will now see morphological operators on functions that act both "horizontally" and "vertically" on them.

As the operators will combine grey-levels by arithmetical additions and subtractions, it will no longer be possible to take a bounded interval for the grey-level set T, otherwise the grey-levels resulting from these operations might overflow out of this interval. Thus T must extend from $-\infty$ to $+\infty$. Let $T' = T \setminus \{-\infty, +\infty\}$; formally we have the following two requirements:

- T is closed under the operations of non-empty numerical supremum and infimum (thus T is a *complete lattice*);
- T' is closed under the operations of addition and subtraction (in other words, T' is a *subgroup* of $\mathbb{R}$).

It is then easily seen that either $T = \overline{\mathbb{R}}$ and $T' = \mathbb{R}$, or there is some $a > 0$ such that $T' = a\mathbb{Z} = \{az \mid z \in \mathbb{Z}\}$ and $T = \overline{a\mathbb{Z}} = a\mathbb{Z} \cup \{-\infty, +\infty\}$; in the second case, we can make a scaling of grey-levels by $1/a$, so here we can suppose without loss of generality that $T = \overline{\mathbb{Z}}$ and $T' = \mathbb{Z}$.

We gave above grey-level analogues of some set-theoretical operations. We have to extend this analogy further. First we redefine the *umbra* or *hypograph* of a function $F : E \to T$, it is the set

$$U(F) = \{(p, t) \in E \times T' \mid t \leq F(p)\} \ .$$

The difference with the previous definition is that we restrict t to T', while before we had $t \in T$. The points (p,t) of the umbra $U(F)$ are the analogues of the points $x \in X$ for a set X. We have now to give the analogue of a singleton, namely a set $\{p\}$ verifying $\{p\} \subseteq X \;\Leftrightarrow\; p \in X$; it is the *impulse* $i_{p,t}$, for $(p,t) \in E \times T'$, defined as follows:

$$\forall x \in E, \qquad i_{p,t}(x) = \begin{cases} t & \text{if } x = p, \\ -\infty & \text{if } x \neq p. \end{cases}$$

We verify indeed that for a function F and an impulse $i_{(p,t)}$, we have $i_{p,t} \leq F \;\Leftrightarrow\; (p,t) \in U(F)$.

We call the *support* of a function F the set

$$\mathrm{supp}(F) = \{p \in E \mid F(p) > -\infty\} \ .$$

Note that $p \in \mathrm{supp}(F)$ iff there exists some $t \in T'$ with $(p,t) \in U(F)$. We will see below that points outside the support are redundant in calculations; in fact, we can assume that F is defined only on its support; conversely if F is defined only on a subset S of E, we extend it to a function on E by setting $F(p) = -\infty$ for all $p \in E \setminus S$.

We defined above the translation of a function by a point. We extend it to the translation by a pair (p,t). Given a function $F : E \to T$ and a pair $(p,t) \in E \times T'$, the *translate of* F *by* (p,t) is the function $F_{(p,t)}$ whose graph is obtained by translating the graph $\{(x, F(x)) \mid x \in E\}$ by p in the first coordinate and by t in the second, that is,

$$\{(y, F_{(p,t)}(y)) \mid p \in E\} = \{(x+p, F(x)+t) \mid x \in E\} \ ,$$

in other words

$$\forall y \in E, \qquad F_{(p,t)}(y) = F(y-p) + t \ .$$

We can now define the Minkowski addition and subtraction of two functions $E \to T$, by analogy with Equation (14.1). Such an analogy already appeared partially in the definition of the dilation and erosion of a function by a set, Equation (14.10), but we have to extend it further. Given two functions $F, G : E \to T$, we define their Minkowski addition $F \oplus G$ and subtraction $F \ominus G$ as follows:

$$\begin{aligned} F \oplus G &= \bigvee_{(p,t) \in U(G)} F_{(p,t)} \ , \\ &= \bigvee_{(p,t) \in U(F)} G_{(p,t)} \ , \\ &= \bigvee \{ i_{(p+p',t+t')} \mid (p,t) \in U(F), (p',t') \in U(G) \} \ ; \\ F \ominus G &= \bigwedge_{(p,t) \in U(G)} F_{(-p,-t)} \ , \\ &= \bigvee \{ i_{(p,t)} \mid (p,t) \in E \times T',\ G_{(p,t)} \leq F \} \ . \end{aligned} \tag{14.12}$$

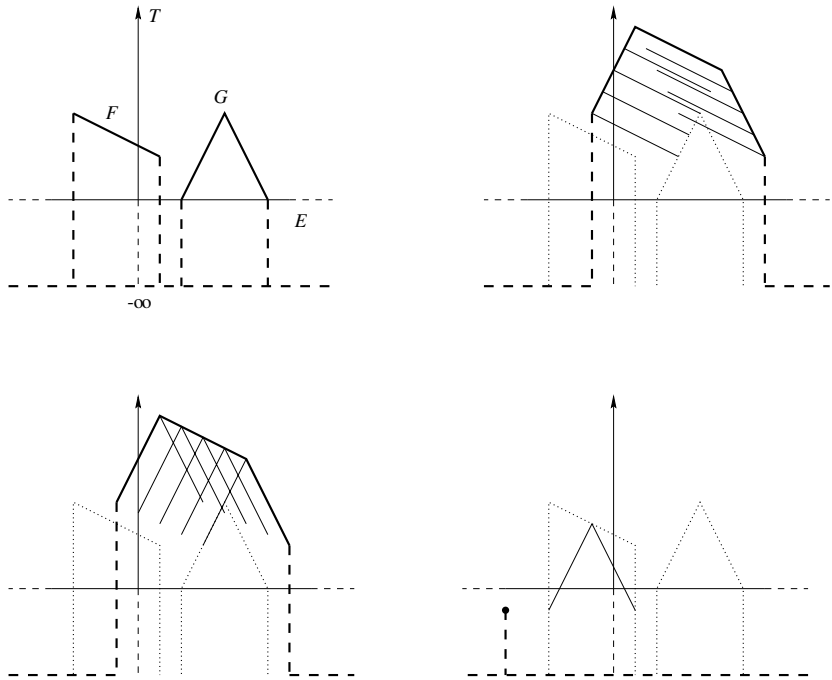


Figure 14.7. Top left: The two grey-level functions F and G, both having value $-\infty$ outside a bounded support. (In the next three illustrations, F and G are shown dotted.) Top right: $F \oplus G$ is the supremum of translates of F by all points of $U(G)$. Bottom left: It is also the supremum of translates of G by all points of $U(F)$. Bottom right: $F \ominus G$ is the supremum of all impulses $i_{(p,t)}$ such that $G_{(p,t)} \leq F$; in fact there is a unique point p for which this is possible, so $F \ominus G$ is an impulse.

These two operations are illustrated in Fig. 14.7. Usually, F plays the role of a grey-level image, while G is the grey-level analogue of a structuring element, and we call it then a *structuring function*.

We can give a numerical expression for the values of $F \oplus G$ and $F \ominus G$; for all $p \in E$ we have

$$
\begin{aligned}
(F \oplus G)(p) &= \sup_{h \in E} \big(F(p-h) + G(h)\big) \\
&= \sup_{h \in \mathrm{supp}(G)} \big(F(p-h) + G(h)\big) \ , \\
(F \ominus G)(p) &= \inf_{h \in E} \big(F(p+h) - G(h)\big) \\
&= \inf_{h \in \mathrm{supp}(G)} \big(F(p+h) - G(h)\big) \ ,
\end{aligned}
\tag{14.13}
$$

with the following convention for dealing with expressions of the form $+\infty - \infty$ inside the parentheses: if $F(p-h) + G(h)$ takes the form $+\infty - \infty$, we set

it equal to $-\infty$, while if $F(p+h) - G(h)$ takes the form $+\infty - \infty$, we set it equal to $+\infty$.

The operators $\delta_G : T^E \to T^E : F \mapsto F \oplus G$ and $\varepsilon_G : T^E \to T^E : F \mapsto F \ominus G$ are called dilation and erosion by G. We can now define the binary operations $\circ$ and $\bullet$ as for sets, by $F \circ G = (F \ominus G) \oplus G$ and $F \bullet G = (F \oplus G) \ominus G$, leading thus to the opening by G, $\gamma_G : T^E \to T^E : F \mapsto F \circ G$, and the closing by G, $\varphi_G : T^E \to T^E : F \mapsto F \bullet G$; note that

$$F \circ G = \bigvee \{G_{(p,t)} \mid (p,t) \in E \times T',\ G_{(p,t)} \leq F\} \ ,$$

which is analogous to the second line in Equation (14.4). We still have the duality by inversion. Define the *transpose* or *symmetrical* $\widetilde{G}$ of G by $\widetilde{G}(x) = G(-x)$; we have the grey-level inversion N on functions (given by $N(F)(p) = -F(p)$); we get then

$$\begin{array}{rclcrcl} N\big(F \oplus G\big) & = & N(F) \ominus \widetilde{G} \ ; & & N\big(F \ominus G\big) & = & N(F) \oplus \widetilde{G} \ ; \\ N\big(F \circ G\big) & = & N(F) \bullet \widetilde{G} \ ; & & N\big(F \bullet G\big) & = & N(F) \circ \widetilde{G} \ . \end{array} \tag{14.14}$$

All properties of these operations $\oplus$, $\ominus$, $\circ$ and $\bullet$ in the case of sets, extend to the case of functions. For example, the opening and closing by G are idempotent. Operators on functions $E \to T$ built from these operations, together with the supremum and infimum, constitute what is called *grey-level morphology* or *functional morphology*.

Note finally that the dilation and erosion of functions by a set structuring element, Eqs. (14.10,14.11), are a particular case of dilation and erosion by a structuring function, Eqs. (14.12,14.13). Given a set $B \subseteq E$, define the function $B_0 : E \to T$ having value 0 on B, and $-\infty$ elsewhere:

$$\forall x \in E, \qquad B_0(x) = \begin{cases} 0 & \text{if } x \in B, \\ -\infty & \text{if } x \notin B. \end{cases}$$

Then for every function $F : E \to T$, we have $F \oplus B = F \oplus B_0$ and $F \ominus B = F \ominus B_0$. The function B_0 is thus called a *flat structuring function*. Hence flat morphology is a particular case of grey-level morphology, with a restriction of structuring functions to flat ones.

We explained above that flat operators behave on bright and dark parts of a grey-level image in the same way as the corresponding set operators do on foreground and background. This remains true here, but now the action is not only on the shape of these parts, but also on their grey-level profiles. For example, dilation and opening deform the grey-level profile on peaks, while erosion and closing do it on valleys. This is illustrated in Fig. 14.8; we see that the opening removes narrow peaks and the closing removes narrow valleys (as expected), but also the slope of jumps is reduced at the top with the opening, and at the bottom with the closing.

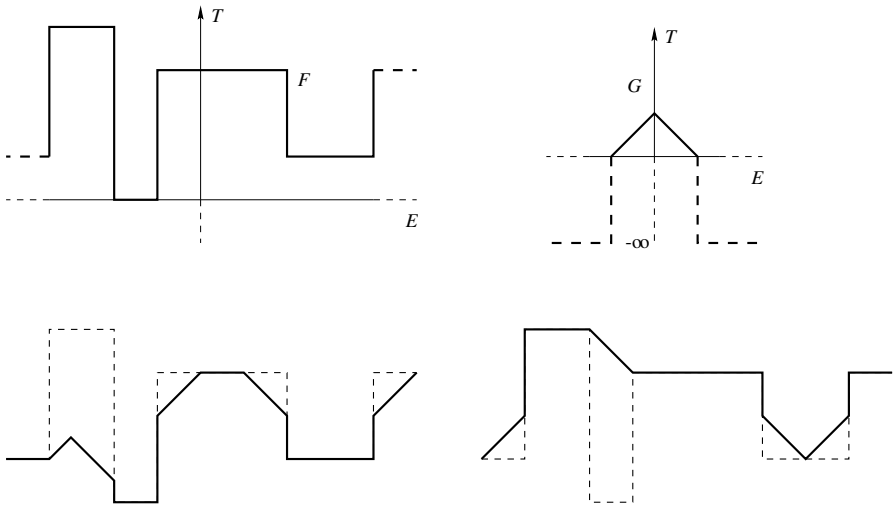


Figure 14.8. Top left: The grey-level function F. Top right: The structuring function G. Bottom left: $F \circ G$, the opening of F (dashed) by G. Bottom right: $F \bullet G$, the closing of F (dashed) by G.

For most practical problems concerning grey-level images, flat morphological operators are applied, instead of functional ones. Indeed, their expression is simpler (as it does not involve adding or subtracting grey-levels), and they work correctly for bounded grey-levels (functional ones can lead to overflow). In fact, flat operators have the same potential as functional ones for dealing with *spatial shapes* of objects in a grey-level image. However, there are sometimes situations where the grey-level profile of objects matters as much as their shape, and in such situations one will use functional morphological operators.

Let us say a few words about the computational complexity of morphological operations. Without any optimization, the complexity of the Minkowski operations is in $O(N \times S)$, where N is the size of the image and S is the size of the structuring element. However, thanks to various approaches, such as the decomposition of structuring elements, or the use of redundancies, it is possible for some particular types of structuring elements (say, rectangles), to have a complexity in $O(N \times \sqrt{S})$, $O(N \times \log S)$, or even $O(N)$. In digital grids using the usual connectivities based on neighbourhoods, geodesical reconstruction has a complexity in $O(N)$, thanks to the use of queues. In the binary case, pixels are inserted in the queue as soon as they receive a connected component label, and leave the queue when they transmit the label to their neighbours. For grey-level reconstruction, one uses a set of queues, one for each grey-level, with a priority order corresponding to the grey-level (e.g., for reconstruction by dilation, priority is given to the highest grey-levels).

2. Algebra

We saw in the Introduction how to define morphological operations on sets by combinations of unions, intersections and translations, and how these operations can be adapted to numerical functions by translating union and intersection into supremum and infimum. For many practical applications, such a framework resting on the analogy between sets and numerical functions, where foreground and background correspond to bright and dark image areas, is sufficient. However, if one wants to deepen the understanding of morphology, two questions come forward:

- Instead of extending the morphology on sets to the one on functions "by analogy", is there not a systematic approach that would give both as particular cases? Indeed, the early studies of grey-level morphology analysed the latter in terms of umbras of functions, they even attempted to make grey-level morphology a particular case of set morphology applied to umbras; however the correspondence between operations on functions and those on umbras is exact only for discrete grey-levels (Ronse, 1990).

- Can we define similarly morphological operations on other types of objects? For example, on the family $\mathcal{F}(\mathbb{R}^n)$ of closed subsets of $\mathbb{R}^n$: here an intersection of closed sets is closed, while a union of closed sets is not closed, but one could take instead the closure of their union; can we adapt Minkowski addition and subtraction in order to obtain all other morphological operators?

The answer to both questions is yes. Morphology on sets, on functions, and on several other types of objects (closed sets, convex sets, etc.) can be seen as particular cases of a general framework based on *complete lattices*. This was first introduced by Serra, 1988, then developed by Heijmans and Ronse, 1990, Ronse and Heijmans, 1991, Heijmans, 1991 and Heijmans, 1994. In this section, we present the algebraic fundamentals of mathematical morphology.

2.1 Complete lattice framework for images and operators

The basic idea is to generalize the notions of inclusion, union and intersection of sets, to other objects.

DEFINITION 14.1 *A* partial order *is a relation* $\leq$ *that is reflexive, anti-symmetrical and transitive. Write* $\geq$ *for the inverse of* $\leq$ *(*$x \geq y$ *iff* $y \leq x$*), it is also a partial order. A* partially ordered set *or* poset *is a pair* $(X, \leq)$*, where* X *is a set and* $\leq$ *a partial order on* X*.*

A complete lattice *is a poset* $(X, \leq)$ *in which every non-void part* Y *of* X *has a* least upper bound *or* supremum $\bigvee Y$*, and a* greatest lower bound *or* infimum $\bigwedge Y$*.*

It follows in particular that a complete lattice $(X, \leq)$ has always a *greatest element*, namely $\bigvee X$, and a *least element*, namely $\bigwedge X$. By analogy with Boolean algebras, the greatest (resp., least) element is also called the *one* (resp., *zero*), and it is written $\mathbf{1}$ or $\top$ (resp., $\mathbf{0}$ or $\perp$). Note also that every $x \in X$ is both lower bound and upper bound of the empty set. Hence:

$$\mathbf{1} = \bigvee X = \bigwedge \emptyset \qquad \text{and} \qquad \mathbf{0} = \bigwedge X = \bigvee \emptyset \ .$$

A *complete sublattice* of X is given by a subset Y of X, such that with the restriction to Y of the order $\leq$ on X, $(Y, \leq)$ is a *complete lattice* in which the supremum and infimum operations, as well as the zero and one, are identical to those in X; equivalently, it is a subset Y of X such that for every $Z \subseteq Y$, $\bigvee Z, \bigwedge Z \in Y$ (also for $Z = \emptyset$, i.e., $\mathbf{0}, \mathbf{1} \in Y$).

Some examples of complete lattices are particularly useful for mathematical morphology:

- The power set $\mathcal{P}(E)$, ordered by the set inclusion; here the supremum and infimum are the union and intersection. It represents the family of binary images.
- The grey-level sets $\overline{\mathbb{R}}$, $\overline{\mathbb{Z}}$, $[a, b]$ and $[a \ldots b]$ considered in Sec. 1.2 are complete lattices, and $\overline{\mathbb{Z}}$ is a complete sublattice of $\overline{\mathbb{R}}$.
- Given T one of the above complete lattices, and a space E, consider the set T^E of numerical functions $E \to T$. It is a complete lattice, in fact a *power lattice* of T, in the sense that its ordering, supremum and infimum derive from those on T by pointwise application, see Eqs. (14.7, 14.8). It represents the family of grey-level images.
- The family $\mathcal{F}(\mathbb{R}^n)$ of closed subsets of $\mathbb{R}^n$ is a complete lattice for the ordering by inclusion; here the infimum of a family of closed sets is its intersection, while its supremum is the closure of its union. Despite the same ordering as in $\mathcal{P}(E)$, it is not a complete sublattice of $\mathcal{P}(E)$, because the supremum operation is not the same. Many metrics and topologies on sets are defined properly only for closed sets (Ronse and Tajine, 2004).
- We can represent RGB colours as triples (r, g, b) of numerical values, so T^3 is the complete lattice of RGB colours, with componentwise ordering; now we represent a RGB colour image as a function $E \to T^3$ associating to each point $p \in E$ a triple $\big(r(p), g(p), b(p)\big)$ coding the RGB colour of p; thus the family of RGB colour images constitutes the complete lattice $(T^3)^E$, with the componentwise ordering.
- Given a set E, the set $\Pi(E)$ of partitions on E is ordered as follows: given two partitions π_1 and π_2, we write $\pi_1 \leq \pi_2$ is π_1 is *finer* than π_2,

or equivalently π_2 is *coarser* than π_1; this means that every class of π_1 is included in a class of π_2. Then $\Pi(E)$ is a complete lattice. In fact, there is a one-to-one correspondence between partitions on E and equivalence relations on E; then the lattice structure of $\Pi(E)$ corresponds by bijection to the one of the family $Equiv(E)$ of equivalence relations on E, considered as a subset of E^2: the ordering on partitions corresponds to the inclusion order between equivalences, and the infimum and supremum of a family of partitions correspond respectively to the intersection and to the transitive closure of the union, of the associated equivalence relations.

Image processing operations can then be viewed as mappings $L \to L$, where L is the complete lattice of images under consideration; we can also consider mappings from one complete lattice to another, for example, $T^E \to \mathcal{P}(E)$ (binarization of grey-level images), or $T^E \to \Pi(E)$ (segmentation). Such mappings are usually written by Greek letters, and are called *operators*; an operator is said *on* L when it is a mapping $L \to L$. Given a set L (which can be a complete lattice or not) and a complete lattice M, the set M^L of operators $L \to M$ is a complete lattice, which inherits the order and complete lattice structure of M "componentwise", as happened for functions, see Eqs. (14.7,14.8): for two operators $\eta, \zeta : L \to M$, we have

$$\eta \leq \zeta \qquad \Longleftrightarrow \qquad \forall x \in L, \quad \eta(x) \leq \zeta(x) \ ,$$

and for a family ψ_i ($i \in I$) of operators $L \to M$, their supremum and infimum are given by:

$$\bigvee_{i \in I} \psi_i : L \to M : x \mapsto \bigvee_{i \in I} \psi_i(x) \quad \text{and} \quad \bigwedge_{i \in I} \psi_i : L \to M : x \mapsto \bigwedge_{i \in I} \psi_i(x) \ .$$

There is another operation on operators, *composition*; given $\eta : L \to M$ and $\zeta : M \to N$, the *composition of* η *followed by* ζ is the operator $\zeta\eta : L \to N : x \mapsto \zeta(\eta(x))$. Of particular interest is the composition of operators on L: the composition of two operators $\zeta, \eta : L \to L$ is always defined, and this gives an associative operation having as neutral element the *identity* $\mathbf{id} : L \to L : x \mapsto x$, in other words the set L^L of operators on L is what one calls a *monoid*. Given an operator ψ on L, we define recursively the power ψ^n for every $n \in \mathbb{N}$: $\psi^0 = \mathbf{id}$, $\psi^{n+1} = \psi[\psi^n]$. Let us recall some morphological terminology (Serra, 1988; Heijmans, 1994):

DEFINITION 14.2 *Given two posets L and M, an operator $\psi : L \to M$ is*

- increasing *(or* isotone*, Birkhoff, 1995) if for all $x, y \in L$, we have $x \leq y \Rightarrow \psi(x) \leq \psi(y)$.*
- decreasing *(or* antitone*, Birkhoff, 1995) if for all $x, y \in L$, we have $x \leq y \Rightarrow \psi(x) \geq \psi(y)$.*

- *an* isomorphism *if ψ is an increasing bijection, whose inverse ψ^{-1} is increasing.*
- *a* dual isomorphism *if ψ is a decreasing bijection, whose inverse ψ^{-1} is decreasing.*

Given a poset L, an operator ψ on L is

- extensive *if $\psi \geq \mathbf{id}$, that is, for every $x \in L$ we have $\psi(x) \geq x$.*
- anti-extensive *if $\psi \leq \mathbf{id}$, that is, for every $x \in L$ we have $\psi(x) \leq x$.*
- *an* automorphism of L *if ψ is an isomorphism $L \to L$.*
- *a* dual automorphism of L *if ψ is a dual isomorphism $L \to L$.*

Given a set L, an operator ψ on L is idempotent *if $\psi\psi = \psi$, that is, for every $x \in L$ we have $\psi(\psi(x)) = \psi(x)$.*

Note that if L is a complete lattice and ψ is an increasing operator on L, then we have (Heijmans and Ronse, 1990):

$$\forall (x_i, i \in I) \subseteq L, \qquad \begin{cases} \psi\Big(\bigvee_{i\in I} x_i\Big) & \geq & \bigvee_{i\in I} \psi(x_i), \\ \psi\Big(\bigwedge_{i\in I} x_i\Big) & \leq & \bigwedge_{i\in I} \psi(x_i). \end{cases} \tag{14.15}$$

Some other properties and specific families of operators (in particular, dilations, erosions, openings and closings) will be defined in the following subsections.

Given an operator ψ on a set L, the *invariance domain of* ψ is the set $Inv(\psi) = \{x \in L \mid \psi(x) = x\}$. Given an operator $\psi : L \to M$, the *range* (or *image*) of ψ is the set of $\psi(x)$ for $x \in L$; we write it $\psi(L)$.

In the lattice $\mathcal{P}(E)$ of parts of a Euclidean or digital space $E = \mathbb{R}^n$ or $\mathbb{Z}^n$, the dilation, erosion, opening and closing by a structuring element, Eqs. (14.1,14.4), are *translation-invariant*, in other words they commute with any translation of E. We can generalize this notion as follows. Let $\mathbf{T}$ be a *group* of automorphisms of the complete lattice L; in other words for every $\tau \in \mathbf{T}$, τ is an automorphism of L and $\tau^{-1} \in \mathbf{T}$, and for every $\tau_1, \tau_2 \in \mathbf{T}$, $\tau_1\tau_2 \in \mathbf{T}$. An operator ψ on L is said to be $\mathbf{T}$-*invariant* if it commutes with every element of $\mathbf{T}$: $\forall \tau \in \mathbf{T}$, $\tau\psi = \psi\tau$.

There is an important principle: *duality*. We saw above that the inverse $\geq$ of a partial order $\leq$ is a partial order. Therefore every notion concerning posets and complete lattices admits a *dual*, which is the same notion expressed w.r.t. the inverse order $\geq$; as the inverse of $\geq$ is again $\leq$, the duality is symmetrical. For example, the dual of the supremum operation is the infimum operation (and vice versa); for an operator, being extensive and being anti-extensive are dual

properties. Note that all notions relying only on composition of operators, and not on order, are *auto-dual*; this is for example the case for the identity operator and for the property of idempotence.

An *inversion* of a poset L is a dual automorphism of L which is its own inverse, in other words a decreasing operator ν on L such that $\nu^2 = \mathbf{id}$. Then every operator ψ on L has a *dual by inversion*, $\psi^* = \nu\psi\nu$, whose properties are dual to those of ψ. For example, the set complementation in $\mathcal{P}(E)$, and the grey-level inversion (image negative) N in T^E, are inversions; in $\mathcal{P}(E)$ the dilation (resp., opening) by B is the dual by complementation of the erosion (resp., closing) by $\check{B}$, Eqs. (14.2,14.5), and in T^E the dilation and opening by G are the duals by inversion of the erosion and closing by $\widetilde{G}$, Equation (14.14).

2.2 Moore families, algebraic closings and openings

There are many mathematical situations where an object is "closed" under some operation: a closed set in a topological space, a convex set in $\mathbb{R}^n$, a subgroup of a group, a transitive relation. The interesting thing is that when an object is not closed, one can close it in a unique smallest possible way. From the algebraic point of view, it is thus fundamental to describe both the structure of the family of closed sets, and the properties of the closure operator.

DEFINITION 14.3 *Let L be a poset.*

1 *A subset M of L is a* Moore family *if every element of L has a least upper bound in M:*

$$\forall x \in L,\ \Big(\exists y \in M, y \geq x \text{ and } \big[\forall z \in M,\ (z \geq x \ \Rightarrow\ z \geq y)\big]\Big)\ .$$

2 *A* closing *(or* closure operator*) on L is an increasing, extensive and idempotent operator $L \to L$.*

The Moore family stands for the family of closed objects. The equivalence between the two concepts of closed object and closing an object, is expressed as follows:

PROPOSITION 14.4 *Let L be a poset. There is a one-to-one correspondence between Moore families in L and closings on L, given as follows:*

- *To a Moore family M we associate the closing φ defined by setting for every $x \in L$: $\varphi(x)$ is equal to the least $y \in M$ such that $y \geq x$.*
- *To a closing φ one associates the Moore family M which is the* invariance domain *of φ: $M = Inv(\varphi)$.*

Note that $M = \{\varphi(x) \mid x \in L\}$. Let us now consider the case where L is a complete lattice.

THEOREM 14.5 *Let L be a complete lattice. A subset M of L is a Moore family iff M is closed under the infimum operation:*

$$\forall S \subseteq M, \ \bigwedge S \in M \ .$$

In particular, $\bigwedge \emptyset = \mathbf{1} \in M$. Given a Moore family M corresponding to a closing φ, $(M, \leq)$ is a complete lattice with greatest element $\mathbf{1}$ and least element $\varphi(\mathbf{0}) = \bigwedge M$, and where the supremum and infimum of a family $N \subseteq M$ are given by $\varphi\big(\bigvee N\big)$ and $\bigwedge N$, respectively.

Note that $\varphi(\mathbf{1}) = \mathbf{1}$ and $\varphi\big(\bigwedge N\big) = \bigwedge N$. Let us mention also the following property:

$$\forall X \subseteq L, \qquad \varphi\Big(\bigvee_{x \in X} \varphi(x)\Big) = \varphi\Big(\bigvee X\Big) \ . \tag{14.16}$$

Let us illustrate the above results with the family $\mathcal{F}$ of closed sets in a topological space E. Clearly $\mathcal{F}$ is a Moore family of $\mathcal{P}(E)$ (ordered by inclusion), which means that $\mathcal{F}$ is closed under arbitrary intersections, and contains the empty intersection $\bigcap \emptyset = E$; now $\mathcal{F}$ corresponds to a closing, which is the topological closure operator $\mathbf{cl}$, where for $X \subseteq E$, $\mathbf{cl}(X)$ is the least element of $\mathcal{F}$ containing X. However the Moore family $\mathcal{F}$ has two further properties:

1 $\emptyset \in \mathcal{F}$; by Theorem 14.5, this is equivalent to $\mathbf{cl}(\emptyset) = \emptyset$.

2 $\mathcal{F}$ is closed under binary union: for $C_1, C_2 \in \mathcal{F}$, $C_1 \cup C_2 \in \mathcal{F}$. By Theorem 14.5, this means that $\mathbf{cl}(C_1 \cup C_2) = C_1 \cup C_2$. Now $\mathcal{F}$ is the set of $\mathbf{cl}(X)$ for $X \in \mathcal{P}(E)$, and we can write $C_i = \mathbf{cl}(X_i)$, so in view of Equation (14.16), the condition is equivalent to:

$$\forall X_1, X_2 \in \mathcal{P}(E), \qquad \mathbf{cl}(X_1 \cup X_2) = \mathbf{cl}(X_1) \cup \mathbf{cl}(X_2) \ .$$

Therefore one can characterize a topology, given by the family of closed sets, through the associated closure operator $\mathbf{cl}$, which must be a closing (increasing, idempotent and extensive), preserve the empty set, and distribute binary union (Everett, 1944).

Let us now consider the dual concepts and results:

- In a poset L, a *dual Moore family* is a subset M such that every element of L has a greatest lower bound in M.
- The dual of a closing is an *opening* on L: an increasing, anti-extensive and idempotent operator $L \to L$.
- There is a one-to-one correspondence between dual Moore families in L and openings on L, where the corresponding opening γ and dual Moore

family M verify: $\gamma(x)$ is equal to the greatest $y \in M$ such that $y \leq x$, and M is the invariance domain of γ.

- In a complete lattice L, M is a dual Moore family iff M is closed under the supremum operation; in particular $\mathbf{0} \in M$. Given a dual Moore family M corresponding to an opening γ, $(M, \leq)$ is a complete lattice with greatest element $\gamma(\mathbf{1}) = \bigvee M$ and least element $\mathbf{0}$, and where the supremum and infimum of a family $N \subseteq M$ are given by $\bigvee N$ and $\gamma\big(\bigwedge N\big)$, respectively.
- A topology on a space E can be characterized by its topological interior operation $\mathbf{int}$, which is an opening verifying $\mathbf{int}(E) = E$ and $\mathbf{int}(X_1 \cap X_2) = \mathbf{int}(X_1) \cap \mathbf{int}(X_2)$.

Let us now describe the structure of the families of openings and closings. This will lead to some standard methods to construct them.

PROPOSITION 14.6 *Let L be a complete lattice.*

1 *The supremum of any family of openings on L is an opening, and the set of openings on L is a dual Moore family in L^L. For every increasing operator ψ on L, the greatest opening $\leq \psi$ is $\Gamma(\psi)$, it verifies $Inv(\Gamma(\psi)) = Inv(\mathbf{id} \wedge \psi)$.*

2 *The infimum of any family of closings is a closing, and the set of closings on L is a Moore family in L^L. For every increasing operator ψ on L, the least closing $\geq \psi$ is $\Phi(\psi)$, it verifies $Inv(\Phi(\psi)) = Inv(\mathbf{id} \vee \psi)$.*

By Proposition 14.4 (and its dual), for any $x \in L$, $\Gamma(\psi)(x)$ is the greatest $y \in Inv(\mathbf{id} \wedge \psi)$ such that $y \leq x$, and $\Phi(\psi)(x)$ is the least $y \in Inv(\mathbf{id} \vee \psi)$ such that $y \geq x$.

By Theorem 14.5 (and its dual), the set of openings (resp., closings) is a complete lattice, where $\mathbf{id}$ is the greatest opening (resp., the least closing), and the least opening is the constant operator $L \to L : x \mapsto \mathbf{0}$ (resp., the greatest closing is the constant operator $L \to L : x \mapsto \mathbf{1}$).

One can construct openings and closings by specifying some of their invariants. Let $b \in L$ and let $\mathbf{T}$ be a group of automorphisms of L. The *structural opening* and *closing* $\gamma_{b,\mathbf{T}}$ and $\varphi_{b,\mathbf{T}}$ are defined by

$$\forall x \in L, \quad \begin{cases} \gamma_{b,\mathbf{T}}(x) = \bigvee\{\tau(b) \mid \tau \in \mathbf{T},\ \tau(b) \leq x\}, \\ \varphi_{b,\mathbf{T}}(x) = \bigwedge\{\tau(b) \mid \tau \in \mathbf{T},\ \tau(b) \geq x\}. \end{cases} \tag{14.17}$$

More generally, given a family $S \subseteq L$, we define then

$$\gamma_{S,\mathbf{T}} = \bigvee_{s \in S} \gamma_{s,\mathbf{T}} \qquad \text{and} \qquad \varphi_{S,\mathbf{T}} = \bigwedge_{s \in S} \varphi_{s,\mathbf{T}} \ , \tag{14.18}$$

and we have

$$(14.19) \quad \forall x \in L, \quad \begin{cases} \gamma_{S,\mathbf{T}}(x) = \bigvee\{\tau(s) \mid s \in S,\ \tau \in \mathbf{T},\ \tau(s) \leq x\}, \\ \varphi_{S,\mathbf{T}}(x) = \bigwedge\{\tau(s) \mid s \in S,\ \tau \in \mathbf{T},\ \tau(s) \geq x\}. \end{cases}$$

These operators are a $\mathbf{T}$-invariant opening and closing, respectively, and in fact every $\mathbf{T}$-invariant opening and closing takes this form:

PROPOSITION 14.7 *Let L be a complete lattice. For any $S \subseteq L$, let $\langle S\rangle_{\mathbf{T}}^{\sup}$ (resp., $\langle S\rangle_{\mathbf{T}}^{\inf}$) be the least subset of L containing S which is closed under $\mathbf{T}$ and under the supremum (resp., infimum) operation. We have*

$$\begin{aligned} \langle S\rangle_{\mathbf{T}}^{\sup} &= \Big\{ \bigvee_{(\tau,s)\in X} \tau(s) \mid X \subseteq \mathbf{T} \times S \Big\} \\ \textit{and} \qquad \langle S\rangle_{\mathbf{T}}^{\inf} &= \Big\{ \bigwedge_{(\tau,s)\in X} \tau(s) \mid X \subseteq \mathbf{T} \times S \Big\}. \end{aligned}$$

Then $\gamma_{S,\mathbf{T}}$ and $\varphi_{S,\mathbf{T}}$ are a $\mathbf{T}$-invariant opening and closing, respectively, with these sets as their respective invariance domain:

$$Inv(\gamma_{S,\mathbf{T}}) = \langle S\rangle_{\mathbf{T}}^{\sup} \qquad \textit{and} \qquad Inv(\varphi_{S,\mathbf{T}}) = \langle S\rangle_{\mathbf{T}}^{\inf} \ .$$

Conversely, every $\mathbf{T}$-invariant opening γ and closing φ take this form: $\gamma = \gamma_{Inv(\gamma),\mathbf{T}}$ and $\varphi = \varphi_{Inv(\varphi),\mathbf{T}}$.

A well-known example is when $L = \mathcal{P}(E)$, for $E = \mathbb{R}^n$ or $\mathbb{Z}^n$, and $\mathbf{T}$ is the group of translations of E. Then the structural opening and closing give the opening and closing by a structuring element: for every $X, B \in \mathcal{P}(E)$ we have $\gamma_{B,\mathbf{T}}(X) = X \circ B$ and $\varphi_{B,\mathbf{T}}(X) = (X^c \circ B^c)^c = X \bullet [\check{B}]^c$ (with $b \in [\check{B}]^c \ \Leftrightarrow\ -b \notin B$). For a family $\mathcal{S}$ of structuring elements, we get the openings and closings of the form given in Equation (14.6). We obtain thus the well-known fact that every translation-invariant opening (resp., closing) is a union of openings (resp., intersection of closings) by structuring elements.

When $\mathbf{T}$ reduces to the identity $\mathbf{id}$, we simply write γ_b, φ_b, γ_S and φ_S. Then the above result characterizes arbitrary openings and closings as being γ_S and φ_S for some $S \subseteq L$.

In the next subsection, we will see how openings and closings arise from dilations and erosions.

2.3 Galois connections and adjunctions

At the beginning of the 19th century, Evariste Galois built a connection between fields of numbers generated by roots of equations, and groups of permutations of these roots. This type of correspondence is the first example of a

general technique used in algebra to build an association between two types of structures. It has thus been named after him.

DEFINITION 14.8 *Let A and B two posets, with two operators $\alpha : B \to A$ and $\beta : A \to B$. We say that α and β form a* Galois connection *if*

$$\forall a \in A,\ \forall b \in B, \qquad a \leq \alpha(b) \iff b \leq \beta(a)\ . \tag{14.20}$$

Note that α and β play symmetrical roles. Galois connections are often used in mathematical morphology to establish a dual isomorphism between two types of structures, thanks to the following result:

PROPOSITION 14.9 *Let A and B two posets, and let $\alpha : B \to A$ and $\beta : A \to B$ form a Galois connection. Then:*

1 *α and β are decreasing, $\alpha = \alpha\beta\alpha$ and $\beta = \beta\alpha\beta$.*

2 *$\alpha\beta$ is a closing on A, $\beta\alpha$ is a closing on B, $Inv(\alpha\beta) = \alpha(B)$ and $Inv(\beta\alpha) = \beta(A)$ (so that $\alpha(B)$ and $\beta(A)$ are Moore families).*

3 *The restriction of β to $\alpha(B)$ is a dual isomorphism $\alpha(B) \to \beta(A)$ whose inverse $\beta(A) \to \alpha(B)$ is the restriction of α to $\beta(A)$.*

One can generally characterize the types of maps α and β which may appear in a Galois connection, but in the case of complete lattices, this characterization is straightforward:

DEFINITION 14.10 *Let A and B be complete lattices. An operator $\alpha : B \to A$ is a* Galois map *if it exchanges supremum and infimum:*

$$\forall (x_i, i \in I) \subseteq B, \qquad \alpha\Big(\bigvee_{i\in I} x_i\Big) = \bigwedge_{i\in I} \alpha(x_i)\ .$$

In particular (for $I = \emptyset$), α maps the least element $\mathbf{0}_B$ of B onto the greatest element $\mathbf{1}_A$ of A.

PROPOSITION 14.11 *Let A and B be complete lattices. Then:*

1 *Given $\alpha : B \to A$ and $\beta : A \to B$ forming a Galois connection, α and β are Galois maps.*

2 *Conversely, given a Galois map $\alpha : B \to A$, there is a unique Galois map $\beta : A \to B$ such that α and β form a Galois connection (and vice versa).*

3 *Given $\alpha_1, \alpha_2 : B \to A$ and $\beta_1, \beta_2 : A \to B$ such that α_i and β_i form a Galois connection for $i = 1, 2$, we have $\alpha_1 \leq \alpha_2 \Leftrightarrow \beta_1 \leq \beta_2$.*

4 *Given $\alpha_i : B \to A$ and $\beta_i : A \to B$ forming a Galois connection for $i \in I$, $\bigwedge_{i\in I} \alpha_i$ and $\bigwedge_{i\in I} \beta_i$ form a Galois connection.*

In other words, Galois maps form a Moore family in the complete lattice of operators $A \to B$ (or $B \to A$), and Galois connection establishes an isomorphism between the two complete lattices of Galois maps $A \to B$ and $B \to A$.

Of particular interest are Galois connections between subsets of two sets, which were characterized by Ore, 1944 in terms of a relation between the points of the two sets:

THEOREM 14.12 *Let V and W two sets.*

1 *Given a relation ρ between elements of V and of W, define*

$$\begin{array}{lll} \alpha_\rho : & \mathcal{P}(W) \to \mathcal{P}(V) : & Y \mapsto \{v \in V \mid \forall w \in Y,\ v\ \rho\ w\}, \\ \beta_\rho : & \mathcal{P}(V) \to \mathcal{P}(W) : & X \mapsto \{w \in W \mid \forall v \in X,\ v\ \rho\ w\}. \end{array}$$

Then α_ρ and β_ρ form a Galois connection.

2 *Conversely, given $\alpha : \mathcal{P}(W) \to \mathcal{P}(V)$ and $\beta : \mathcal{P}(V) \to \mathcal{P}(W)$ forming a Galois connection, there is a unique relation ρ between elements of V and of W, such that $\alpha = \alpha_\rho$ and $\beta = \beta_\rho$; the relation ρ is given by*

$$\forall v \in V, \forall w \in W, \qquad v\ \rho\ w \iff v \in \alpha(\{w\}) \iff w \in \beta(\{v\}).$$

Following Birkhoff, 1995, the Galois maps α_ρ and β_ρ are called *polarities*. Galois connections between sets expressed in such a form, arise in many aspects of mathematics and computer science. See for example Sec. 3.1.

We turn now to the notion of adjunction, which is "semi-dual" to the one of Galois connection, in the sense that we reverse the ordering on one of the posets, but not on the other.

DEFINITION 14.13 *Let A and B two posets, with two operators and $\delta : A \to B$ and $\varepsilon : B \to A$. We say that (ε, δ) is an* adjunction *if*

$$\forall a \in A,\ \forall b \in B, \qquad \delta(a) \leq b \iff a \leq \varepsilon(b)\ . \tag{14.21}$$

We say that δ is lower adjoint *of ε, and ε is* upper adjoint *of δ.*

Compared with Galois connections (see Definition 14.8), we have reversed the ordering on B, since we have $\delta(a) \leq b$ instead of $b \leq \delta(a)$. Hence ε and δ do not play symmetrical roles, that is why we write the ordered pair (ε, δ). We obtain then the analogue of Proposition 14.9:

PROPOSITION 14.14 *Let A and B two posets, and let $\delta : A \to B$ and $\varepsilon : B \to A$ such that (ε, δ) is an adjunction. Then:*

1 *ε and δ are increasing, $\varepsilon = \varepsilon\delta\varepsilon$ and $\delta = \delta\varepsilon\delta$.*

2 *$\varepsilon\delta$ is a closing on A, $\delta\varepsilon$ is an opening on B, $Inv(\varepsilon\delta) = \varepsilon(B)$ and $Inv(\delta\varepsilon) = \delta(A)$ (so that $\varepsilon(B)$ is a Moore family and $\delta(A)$ is a dual Moore family).*

3 *The restriction of δ to $\varepsilon(B)$ is an isomorphism $\varepsilon(B) \to \delta(A)$ whose inverse $\delta(A) \to \varepsilon(B)$ is the restriction of ε to $\delta(A)$.*

PROPOSITION 14.15 *Let L be a poset, $\mathbf{T}$ a group of automorphisms of L, and $\varepsilon, \delta : L \to L$ such that (ε, δ) is an adjunction. Then ε is $\mathbf{T}$-invariant iff δ is $\mathbf{T}$-invariant.*

Let us now characterize adjunctions in the case of complete lattices.

DEFINITION 14.16 *Let A and B be complete lattices.*

1 *An operator $\varepsilon : B \to A$ is an* erosion *if it commutes with the infimum operation:*

$$\forall (x_i, i \in I) \subseteq B, \qquad \varepsilon\Big(\bigwedge_{i\in I} x_i\Big) = \bigwedge_{i\in I} \varepsilon(x_i) \ .$$

In particular (for $I = \emptyset$), ε maps the greatest element $\mathbf{1}_B$ of B onto the greatest element $\mathbf{1}_A$ of A.

2 *An operator $\delta : B \to A$ is a* dilation *if it commutes with the supremum operation:*

$$\forall (x_i, i \in I) \subseteq B, \qquad \delta\Big(\bigvee_{i\in I} x_i\Big) = \bigvee_{i\in I} \delta(x_i) \ .$$

In particular (for $I = \emptyset$), δ maps the least element $\mathbf{0}_B$ of B onto the least element $\mathbf{0}_A$ of A.

Note that dilations and erosions are increasing. Also the set of $\delta(x)$ ($x \in B$) is closed under the supremum operation, while the set of $\varepsilon(x)$ ($x \in B$) is closed under the infimum operation. We obtain now the analogue of Proposition 14.11:

THEOREM 14.17 *Let A and B be complete lattices. Then:*

1 *Given $\delta : A \to B$ and $\varepsilon : B \to A$ such that (ε, δ) is an adjunction, δ is a dilation and ε is an erosion.*

2 *Conversely, (a) given a dilation $\delta : A \to B$, there is a unique erosion $\varepsilon : B \to A$ such that (ε, δ) is an adjunction, and*

(*b*) *given an erosion* $\varepsilon : B \to A$*, there is a unique dilation* $\delta : A \to B$ *such that* (ε, δ) *is an adjunction.*

3 *Given* $\delta_1, \delta_2 : A \to B$ *and* $\varepsilon_1, \varepsilon_2 : B \to A$ *such that* $(\varepsilon_i, \delta_i)$ *is an adjunction for* $i = 1, 2$*, we have* $\delta_1 \leq \delta_2 \;\Leftrightarrow\; \varepsilon_1 \geq \varepsilon_2$.

4 *Given* $\delta_i : A \to B$ *and* $\varepsilon_i : B \to A$ *such that* $(\varepsilon_i, \delta_i)$ *is an adjunction for* $i \in I$*,* $\left(\bigwedge_{i\in I} \varepsilon_i, \bigvee_{i\in I} \delta_i\right)$ *is an adjunction.*

In other words, in the complete lattice of operators $A \to B$ *(or* $B \to A$*), erosions form a Moore family, while dilations form a dual Moore family, and adjunctions establish a dual isomorphism between the two complete lattices of dilations* $A \to B$ *and erosions* $B \to A$.

The classical example of adjunction is given by the erosion and dilation by a structuring element or function, Eqs. (14.1, 14.10, 14.12), arising from the Minkowski addition and subtraction. They are both translation-invariant (cf. Proposition 14.15). Here $A = B = \mathcal{P}(E)$ or T^E. In fact, every translation invariant dilation/erosion on sets arises from Minkowski operations, Equation (14.1), while for functions, every flat dilation/erosion invariant under spatial translations takes the form of Equation (14.10), and every dilation/erosion invariant under both spatial and grey-level translations arises from Minkowski operations, Equation (14.12).

In Heijmans and Ronse, 1990, there is a general study of complete lattices where it is possible to define such Minkowski operations, and to obtain for them properties similar to those verified for sets. Particular cases include of courses $\mathcal{P}(E)$ and T^E ($E = \mathbb{R}^n$ or $\mathbb{Z}^n$, $T = \overline{\mathbb{R}}$ or $\overline{\mathbb{Z}}$), for which we obtain the form given in Eqs. (14.1, 14.12), but also: the lattice of convex subsets of $\mathbb{R}^n$ (here the supremum is the convex hull of the union, but Minkowski operations are the same as in $\mathcal{P}(\mathbb{R}^n)$), the lattice $\mathcal{F}(\mathbb{R}^n)$ of closed sets of $\mathbb{R}^n$ (here the supremum is the closure of the union, and the Minkowski addition is the closure of the one obtained in $\mathcal{P}(\mathbb{R}^n)$, but the Minkowski subtraction is the one of $\mathcal{P}(\mathbb{R}^n)$), upper semi-continuous functions $R^n \to \overline{\mathbb{R}}$, etc.

In the case where $A = B$, the operators ε, δ are $A \to A$, and can be composed arbitrarily in any order. It is then easily checked that in a poset A we have

$$\delta \geq \mathbf{id} \;\Leftrightarrow\; \delta \geq \varepsilon\delta \;\Leftrightarrow\; \delta\varepsilon \geq \varepsilon \;\Leftrightarrow\; \mathbf{id} \geq \varepsilon \tag{14.22}$$

and

$$\delta^2\varepsilon \leq \mathbf{id} \;\Leftrightarrow\; \delta^2 \leq \delta \;\Leftrightarrow\; \delta \leq \varepsilon\delta \;\Leftrightarrow\; \delta\varepsilon \leq \varepsilon \;\Leftrightarrow\; \varepsilon \leq \varepsilon^2 \;\Leftrightarrow\; \mathbf{id} \leq \varepsilon^2\delta \;. \tag{14.23}$$

This gives then the following result, which will be used later on, in the case of sets:

PROPOSITION 14.18 *Let A be a poset, and let (ε, δ) be an adjunction (for $\delta, \varepsilon : A \to A$). Then the following five statement are equivalent: (a) δ is a closing, (b) ε is an opening, (c) $\delta\varepsilon = \varepsilon$, (d) $\varepsilon\delta = \delta$, (e) δ and ε verify one statement of Equation (14.22) and one statement of Equation (14.23). Then we have*

$$\begin{aligned} &Inv(\varepsilon\delta) = Inv(\delta\varepsilon) = Inv(\delta) = Inv(\varepsilon) \\ &= \varepsilon\delta(A) = \delta\varepsilon(A) = \delta(A) = \varepsilon(A) \ . \end{aligned}$$

This set is both a Moore family and a dual Moore family in A; when A is a complete lattice, it is a complete sublattice of A.

Let us now consider dilations, erosions and adjunctions on sets. Let V and W two sets, and let ρ be a relation between elements of V and of W. We define $\delta_\rho : \mathcal{P}(V) \to \mathcal{P}(W)$, the dilation by ρ, and $\varepsilon_\rho : \mathcal{P}(W) \to \mathcal{P}(V)$, the erosion by ρ, as follows:

$$\begin{aligned} &\forall X \in \mathcal{P}(V), \quad \delta_\rho(X) = \{w \in W \mid \exists v \in X,\ v\ \rho\ w\}, \\ &\forall Y \in \mathcal{P}(W), \quad \varepsilon_\rho(Y) = \{v \in V \mid \forall w \in W,\ v\ \rho\ w \Rightarrow w \in Y\}. \end{aligned} \tag{14.24}$$

Alternately, we can define dilation erosion in terms of a map $N : V \to \mathcal{P}(W)$ and the *dual* map $\widetilde{N} : W \to \mathcal{P}(V)$, corresponding to the relation ρ by

$$\forall v \in V, \forall w \in W, \quad \left\{ \begin{array}{rl} & w \in N(v) \ \Leftrightarrow\ v \in \widetilde{N}(w) \ \Leftrightarrow\ v\ \rho\ w, \\ \text{that is,} & N(v) = \{w \in W \mid v\ \rho\ w\}. \\ \text{and} & \widetilde{N}(w) = \{v \in V \mid v\ \rho\ w\}. \end{array} \right. \tag{14.25}$$

When $V = W$, the set $N(v)$ can be considered as the window or neighbourhood of point v, and N is called a *neighbourhood function* or a *windowing function*. Now Equation (14.24) can be written

(14.26)

$$\begin{aligned} &\forall X \in \mathcal{P}(V), \quad \delta_N(X) = \bigcup_{v \in X} N(v) = \{w \in W \mid \widetilde{N}(w) \cap X \neq \emptyset\} \ , \\ &\forall Y \in \mathcal{P}(W), \quad \varepsilon_N(Y) = \{v \in V \mid N(v) \subseteq Y\} \ . \end{aligned}$$

We have then the analogue for adjunctions of Ore's characterization of Galois connections on sets (Theorem 14.12):

THEOREM 14.19 *Let V and W two sets.*

1 *Given a map $N : V \to \mathcal{P}(W)$, $(\varepsilon_N, \delta_N)$ is an adjunction.*

2 *Conversely, given $\delta : \mathcal{P}(V) \to \mathcal{P}(W)$ and $\varepsilon : \mathcal{P}(W) \to \mathcal{P}(V)$ such that (ε, δ) is an adjunction, there is a unique map $N : V \to \mathcal{P}(W)$ such that $\delta = \delta_N$ and $\varepsilon = \varepsilon_N$; for every $v \in V$, $N(v) = \delta(\{v\})$.*

Note that $\delta_{\widetilde{N}}$ is a dilation $\mathcal{P}(W) \to \mathcal{P}(V)$, $\varepsilon_{\widetilde{N}}$ is an erosion $\mathcal{P}(V) \to \mathcal{P}(W)$, $(\varepsilon_{\widetilde{N}}, \delta_{\widetilde{N}})$ is an adjunction, and that $\delta_{\widetilde{N}}$ and $\varepsilon_{\widetilde{N}}$ are dual by complementation of ε_N and δ_N respectively, as

$$\begin{array}{rll} & \forall Y \in \mathcal{P}(W), & \delta_{\widetilde{N}}(Y) = V \setminus \varepsilon_N(W \setminus Y) \\ \text{and} & \forall X \in \mathcal{P}(V), & \varepsilon_{\widetilde{N}}(X) = W \setminus \delta_N(V \setminus X). \end{array}$$

In fact $\delta_{\widetilde{N}} = \delta_{\rho^{-1}}$ and $\varepsilon_{\widetilde{N}} = \varepsilon_{\rho^{-1}}$, where ρ^{-1} is the relation inverse of ρ ($w \ \rho^{-1} \ v \ \Leftrightarrow \ v \ \rho \ w$).

A classical example is given for $V = W = E$ for E being the Euclidean space $\mathbb{R}^n$ or the digital space $\mathbb{Z}^n$, and the neighbourhoods being built from a structuring element $B \subseteq E$: for every $p \in E$, $N(p) = B_p$. Then $\widetilde{N}(p) = (\check{B})_p$ for all $p \in E$, $\delta_N = \delta_B$, $\varepsilon_N = \varepsilon_B$, $\delta_{\widetilde{N}} = \delta_{\check{B}}$ and $\varepsilon_{\widetilde{N}} = \varepsilon_{\check{B}}$. These operators are translation-invariant. In fact, from Proposition 14.15, for an adjunction $(\varepsilon_N, \delta_N)$, ε_N is translation-invariant iff δ_N is translation-invariant, and in such a case it is easily seen that they are the erosion and dilation by the structuring element $B = N(o)$.

PROPOSITION 14.20 *The following are equivalent:*

$$\big(\forall v \in V,\ N(v) \neq \emptyset\big) \ \Leftrightarrow \ \varepsilon_N(\emptyset) = \emptyset \ \Leftrightarrow \ \delta_{\widetilde{N}}(W) = V \ \Leftrightarrow \ \varepsilon_N \leq \delta_{\widetilde{N}} \ .$$

Dually, the following are equivalent:

$$\big(\forall w \in W,\ \widetilde{N}(w) \neq \emptyset\big) \ \Leftrightarrow \ \varepsilon_{\widetilde{N}}(\emptyset) = \emptyset \ \Leftrightarrow \ \delta_N(V) = W \ \Leftrightarrow \ \varepsilon_{\widetilde{N}} \leq \delta_N \ .$$

This result will intervene later on, in particular in Sec. 3.2 and Sec. 4. Note that in the case where $V = W = E$ ($E = \mathbb{R}^n$ or $\mathbb{Z}^n$) and $N(p) = B_p$ for all $p \in E$, the two equivalences reduce both to $B \neq \emptyset$.

Consider now the case where $V = W = E$. Here ρ is a relation on E, and both N and $\widetilde{N}$ are $E \to \mathcal{P}(E)$. The following two results will be used in Sec. 3.2:

PROPOSITION 14.21 *Consider a relation ρ on a set E, and the corresponding maps $N, \widetilde{N} : E \to \mathcal{P}(E)$. Then:*

1 *The following five statements are equivalent: (a) ρ is reflexive, (b) δ_N is extensive, (c) ε_N is anti-extensive, (d) $\delta_{\widetilde{N}}$ is extensive, (e) $\varepsilon_{\widetilde{N}}$ is anti-extensive.*

2 *The following five statements are equivalent: (a) ρ is symmetrical, (b) $\varepsilon_{\widetilde{N}}\delta_N$ is extensive, (c) $\delta_N\varepsilon_{\widetilde{N}}$ is anti-extensive, (d) $\varepsilon_N\delta_{\widetilde{N}}$ is extensive, (e) $\delta_{\widetilde{N}}\varepsilon_N$ is anti-extensive.*

3 *The following five statements are equivalent: (a) ρ is transitive, (b) $\delta_N^2 \leq \delta_N$, (c) $\varepsilon_N^2 \geq \varepsilon_N$, (d) $\delta_{\widetilde{N}}^2 \leq \delta_{\widetilde{N}}$, (e) $\varepsilon_{\widetilde{N}}^2 \geq \varepsilon_{\widetilde{N}}$.*

Combining items 1 and 3 with Proposition 14.18, we deduce:

PROPOSITION 14.22 *Consider a relation ρ on E, and the corresponding maps $N, \widetilde{N} : E \to \mathcal{P}(E)$. Then the following nine statements are equivalent: (a) ρ is reflexive and transitive, (b) δ_N is a closing, (c) ε_N is an opening, (d) $\delta_N\varepsilon_N = \varepsilon_N$, (e) $\varepsilon_N\delta_N = \delta_N$, (f) $\delta_{\widetilde{N}}$ is a closing, (g) $\varepsilon_{\widetilde{N}}$ is an opening, (h) $\delta_{\widetilde{N}}\varepsilon_{\widetilde{N}} = \varepsilon_{\widetilde{N}}$, (i) $\varepsilon_{\widetilde{N}}\delta_{\widetilde{N}} = \delta_{\widetilde{N}}$. We have then*

$$\begin{aligned} &Inv(\varepsilon_N\delta_N) = Inv(\delta_N\varepsilon_N) = Inv(\delta_N) = Inv(\varepsilon_N) \\ &= \{\varepsilon_N\delta_N(Z) \mid Z \in \mathcal{P}(E)\} = \{\delta_N\varepsilon_N(Z) \mid Z \in \mathcal{P}(E)\} \\ &= \{\delta_N(Z) \mid Z \in \mathcal{P}(E)\} = \{\varepsilon_N(Z) \mid Z \in \mathcal{P}(E)\} \ , \end{aligned}$$

and the same with $\widetilde{N}$ in place of N. The two families $Inv(\varepsilon_N\delta_N)$=$Inv(\delta_N\varepsilon_N)$ and $Inv(\varepsilon_{\widetilde{N}}\delta_{\widetilde{N}})$=$Inv(\delta_{\widetilde{N}}\varepsilon_{\widetilde{N}})$ are closed under arbitrary union and intersection, and contain E and $\emptyset$ (in other words they are complete sublattices of $(\mathcal{P}(E), \subseteq)$).

An *Alexandroff topology* (Alexandroff, 1937; Alexandroff and Hopf, 1935) is a topological space $(E, \mathcal{G})$ where the family $\mathcal{G}$ of open sets is closed under arbitrary intersection; in other words $\mathcal{G}$ is a complete sublattice of $(\mathcal{P}(E), \subseteq)$. It is equivalent to require that every point of E has a least open neighbourhood. By the *Alexandroff specialization theorem* (Alexandroff, 1956), there is a one-to-one correspondence between Alexandroff topologies on E and reflexive and transitive relations on E; in fact, for $x, y \in E$, $x \ \rho \ y$ iff x is in the closure of $\{y\}$, i.e., iff y belongs to the least neighbourhood of x. It follows then that for $x \in E$, $N(x)$ is the least neighbourhood of x and $\widetilde{N}(x)$ is the topological closure of $\{x\}$, while for $X \in \mathcal{P}(E)$, $\delta_N(X)$ is the least open set containing X (called the *star* of X), $\varepsilon_N(X)$ is the topological interior of X, $\delta_{\widetilde{N}}(X)$ is the topological closure of X, and $\varepsilon_{\widetilde{N}}(X)$ is the greatest closed subset of X. Note that $Inv(\varepsilon_N\delta_N) = Inv(\delta_N\varepsilon_N)$ is the family of open sets and $Inv(\varepsilon_{\widetilde{N}}\delta_{\widetilde{N}}) = Inv(\delta_{\widetilde{N}}\varepsilon_{\widetilde{N}})$ is the family of closed sets.

We saw in Sec. 2.2 that a closing φ on $\mathcal{P}(E)$ is the closure operator in a topology on E iff it satisfies the following two additional constraints: $\varphi(\emptyset) = \emptyset$ and $\varphi(X_1 \cup X_2) = \varphi(X_1) \cup \varphi(X_2)$ for all $X_1, X_2 \in \mathcal{P}(E)$; we have then $\varphi(X_1 \cup \cdots \cup X_n) = \varphi(X_1) \cup \cdots \cup \varphi(X_n)$ for all $X_1, \ldots, X_n \in \mathcal{P}(E)$. In other words the commutation with the union operation, $\varphi\big(\bigcup_{i \in I} X_i\big) = \bigcup_{i \in I} \varphi(X_i)$, is verified for I being empty or finite. This is weaker than φ being a dilation, where this identity is verified also for an infinite family I; but then the set of closed sets $\varphi(X)$ is closed under infinite unions, which means indeed that we have an Alexandroff topology.

2.4 Morphological filters

The word "filter" is used in several scientific and technological contexts, with various meanings. In image processing, one knows the linear filters, namely

convolution operators, in particular the bandpass filter from signal processing, which preserves all frequencies within a band, and eliminates all others. In non-linear image processing, the well-known median filter has been used to remove impulsive noise, without the blurring effect of linear smoothing filters. The morphological approach to filtering is similar to that of signal processing, namely preserving some parts of an image and eliminating some others, except that the separation of these parts is not based on frequencies. The model proposed is that of an *ideal filter*, i.e., one that keeps the wanted components unaltered, and eliminates completely the unwanted ones. In order to characterize an ideal filter, rather than describing the features to be preserved or removed, one takes an algebraic point of view: if the filter does not alter the wanted parts and eliminates completely the unwanted ones, then applying the filter a second time will not change anything. Hence the main characteristic of an ideal filter is its idempotence. This is important from a theoretical point of view, but also for practical applications: if after applying the filter on an image the result is not satisfying, then we know that another filter must be applied. This contrasts with the behaviour of the median filter: after one application, some noise remains, that could be eliminated by a second or third application; then one can repeat the application of the filter, without guarantee that this will lead to a stable final result, as the median filter can produce oscillations (Serra, 1988). This is related to the fact that one cannot characterize precisely what are the features preserved or eliminated by this filter.

Besides idempotence, mathematical morphology demands that the behaviour of a filter should be related to the order and complete lattice structure of the family of images. Therefore one calls a *morphological filter* (or simply, a *filter*) an increasing and idempotent operator on a poset (or complete lattice). Write $Filt(L)$ for the set of filters on L. We have already encountered some filters: openings and closings. There are many other ones, and we will describe here some techniques for constructing them. This requires some terminology:

DEFINITION 14.23 *Let L be a poset and ψ an operator on L. We say that:*

1. *ψ is* underpotent *if $\psi^2 \leq \psi$.*
2. *ψ is* overpotent *if $\psi^2 \geq \psi$.*
3. *ψ is an* underfilter *if ψ is increasing and underpotent.*
4. *ψ is an* overfilter *if ψ is increasing and overpotent.*

We saw in Proposition 14.6 that in a complete lattice, the set of openings is a dual Moore family and the set of closings is a Moore family. They constitute thus two complete lattices. We have a similar result for filters (Serra, 1988):

PROPOSITION 14.24 *Let L be a complete lattice.*

1 *The set of overfilters on L is a dual Moore family in L^L, i.e. it is closed under the supremum operation.*

2 *The set of underfilters on L is a Moore family in L^L, i.e. it is closed under the infimum operation.*

3 *The set $Filt(L)$ of filters on L is a complete lattice. For any family ψ_i ($i \in I$) of filters, their supremum in $Filt(L)$ is the least underfilter ψ such that $\psi \geq \bigvee_{i\in I} \psi_i$, and their infimum in $Filt(L)$ is the greatest overfilter ψ such that $\psi \leq \bigwedge_{i\in I} \psi_i$.*

This gives a first method for constructing a filter from a family of filters. The second one arises from composition (Serra, 1988):

PROPOSITION 14.25 *Let L be a complete lattice and let ξ and ψ be two filters on L such that $\xi \geq \psi$. Then:*

1 *The only operators that can be obtained by repeated compositions of ψ and ξ are $\psi\xi$, $\xi\psi$, $\psi\xi\psi$ and $\xi\psi\xi$. They are all filters and*

$$\xi \geq \xi\psi\xi \geq \left\{ \begin{array}{c} \psi\xi \\ \xi\psi \end{array} \right\} \geq \psi\xi\psi \geq \psi.$$

2 $Inv(\xi) \cap Inv(\psi) \subseteq Inv(\xi\psi\xi) = Inv(\xi\psi) \subseteq Inv(\xi)$ *and* $Inv(\xi) \cap Inv(\psi) \subseteq Inv(\psi\xi\psi) = Inv(\psi\xi) \subseteq Inv(\psi)$.

3 *In $Filt(L)$, the supremum and infimum of $\xi\psi$ and $\psi\xi$ are $\xi\psi\xi$ and $\psi\xi\psi$ respectively.*

Note that items 1 and 2 do not require L to be a complete lattice, they are valid in any poset. A classical example is when ξ is a closing and ψ is an opening: the opening filters out positive noise, the closing filters out negative noise, so the composition of the two should filter out both types of noise (cf. Sec. 1.1).

The above result is at the basis of a well-known filter introduced in the 1980s, the *alternate sequential filter*. Suppose that we have an image where features of foreground and background are imbricated. To extract an object of a given size, it is necessary to filter its holes at a smaller size, and this require filtering objects at an even smaller size, etc. Thus we will apply openings and closings at increasing scales in order to simplify the image. Consider n openings $\gamma_1, \ldots, \gamma_n$ such that $\gamma_n \leq \cdots \leq \gamma_1$, and n closings $\varphi_1, \ldots, \varphi_n$ such that $\varphi_1 \leq \cdots \leq \varphi_n$. From the previous proposition, the compositions $\mu_i = \gamma_i\varphi_i$, $\nu_i = \varphi_i\gamma_i$, $\rho_i = \varphi_i\gamma_i\varphi_i$ and $\sigma_i = \gamma_i\varphi_i\gamma_i$ are filters. Alternate sequential filters are then defined as:

$$\begin{aligned} \mu_i\mu_{i-1}\cdots\mu_2\mu_1 &= (\gamma_i\varphi_i)(\gamma_{i-1}\varphi_{i-1})\cdots(\gamma_2\varphi_2)(\gamma_1\varphi_1), \\ \nu_i\nu_{i-1}\cdots\nu_2\nu_1 &= (\varphi_i\gamma_i)(\varphi_{i-1}\gamma_{i-1})\cdots(\varphi_2\gamma_2)(\varphi_1\gamma_1), \end{aligned} \tag{14.27}$$

or as the following variants:
(14.28)

$$\begin{aligned}
\rho_i\rho_{i-1}\cdots\rho_2\rho_1 &= (\varphi_i\gamma_i\varphi_i)(\varphi_{i-1}\gamma_{i-1}\varphi_{i-1})\cdots(\varphi_2\gamma_2\varphi_2)(\varphi_1\gamma_1\varphi_1)\\
&= \varphi_i\mu_i\mu_{i-1}\cdots\mu_2\mu_1 \ ,\\
\sigma_i\sigma_{i-1}\cdots\sigma_2\sigma_1 &= (\gamma_i\varphi_i\gamma_i)(\gamma_{i-1}\varphi_{i-1}\gamma_{i-1})\cdots(\gamma_2\varphi_2\gamma_2)(\gamma_1\varphi_1\gamma_1)\\
&= \gamma_i\nu_i\nu_{i-1}\cdots\nu_2\nu_1 \ ,
\end{aligned}$$

for $i = 1, \ldots, n$. They are all filters. They are useful for filtering images where grains (bright zones) are imbricated with pores (dark zones) at all sizes. Typically, the γ_i's and φ_i's can be:

- openings and closings by structuring elements of increasing sizes;
- openings and closings by reconstruction, based on structuring elements of increasing sizes;
- area openings and closings (removing grains and pores on the basis of their area), with increasing area thresholds;

hence as i increases, the alternating sequential filters will progressively remove grains and pores of increasing sizes, thus simplifying the image. (We will discuss further the notion of removing "features of increasing sizes" in Sec. 2.5.) An example is provided in Fig. 14.9 and 14.10.

Schonfeld and Goutsias, 1991 noticed that besides Eqs. (14.27,14.28), *any* composition of openings γ_i and closings φ_i, *in any order*, is a filter. Their argument was generalized by Heijmans, 1997 as follows:

PROPOSITION 14.26 *Let $\psi_1, \ldots, \psi_n$ be overfilters and $\xi_1, \ldots, \xi_n$ be underfilters such that*

$$\psi_n \leq \cdots \leq \psi_1 \leq \xi_1 \leq \cdots \xi_n \ .$$

Then any composition of these operators, containing at least one ψ_i and one ξ_j, is a filter.

A consequence is the following surprising result:

PROPOSITION 14.27 *Let (ε, δ) be an adjunction in a poset L. Then any repeated composition of ε and δ in any order, containing the same number of instances of ε and of δ, is a filter.*

More precisely, an operator of the form $\psi_1 \cdots \psi_{2n}$, where for each $i = 1, \ldots, 2n$ we have $\psi_i \in \{\delta, \varepsilon\}$, and $card\{i = 1, \ldots, 2n \mid \psi_i = \delta\} = card\{i = 1, \ldots, 2n \mid \psi_i = \varepsilon\}$, is a filter.

For more results on filters, the reader is referred to Heijmans, 1994, Heijmans, 1997, Ronse and Heijmans, 1991, Serra, 1988 and Soille, 2003.

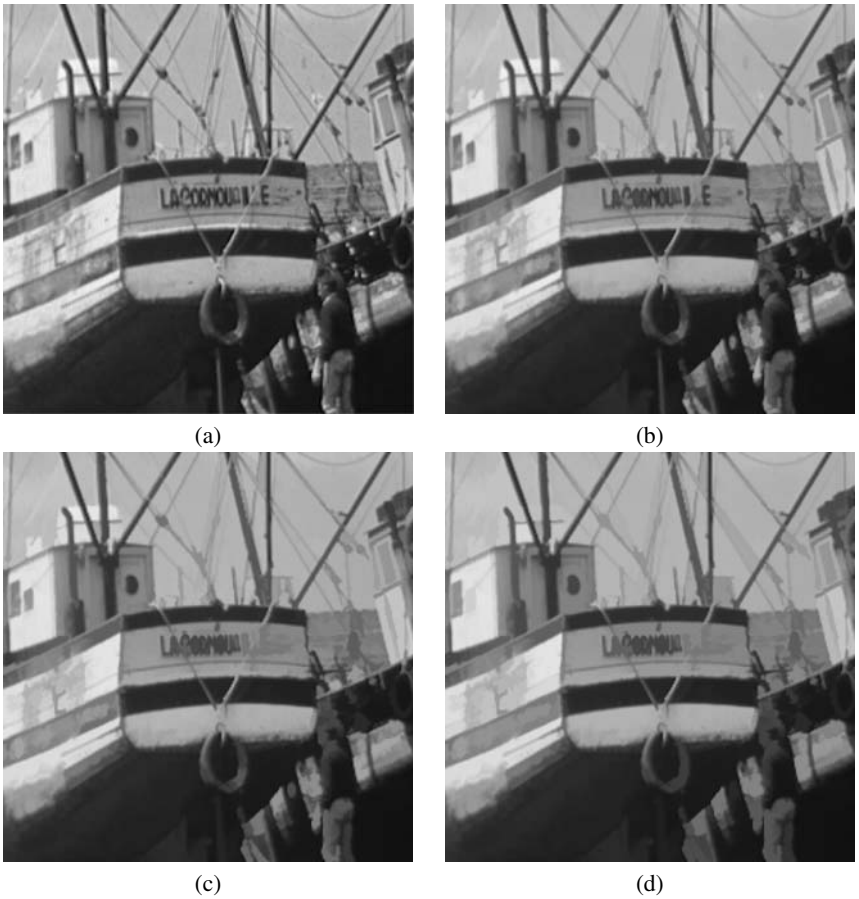


Figure 14.9. Original image (a) and three steps (b, c, d) of an alternate sequential filter based on opening-closing by reconstruction using an hexagon as structuring element.

2.5 Granulometries and size distributions

As openings remove parts of an object (they are anti-extensive), one can compare two openings γ_1 and γ_2 in such terms; thus we say that γ_2 is *more active* than γ_1 if γ_2 removes from any object more than γ_1 does, in other words if $\gamma_2 \leq \gamma_1$. On the other hand, as closings add parts to an object, given two closings φ_1 and φ_2, we say that φ_2 is *more active* than φ_1 if φ_2 adds to an object more than φ_1 does, in other words if $\varphi_2 \geq \varphi_1$.

In the case of the complete lattice $\mathcal{P}(E)$, given two structuring elements $A, B \in \mathcal{P}(E)$, we define the relation $\sqsupseteq$ by $B \sqsupseteq A$ iff B is a union of translates

(a) (b)

Figure 14.10. Steps 1 (a) and 3 (b) of the same filter as in Fig. 14.9 but using segments in different directions as structuring elements.

of A. Readily, by Eqs. (14.1,14.4) we have

$$B \sqsupseteq A \iff B \circ A = B \iff \big(\exists C \in \mathcal{P}(E),\ B = A \oplus C\big) \ .$$

For example, given $b \geq a$, this is true if A and B are squares of size a and b respectively, or (for $E = \mathbb{R}^n$) if A and B are closed balls of radii a and b respectively. Now by Equation (14.4), the openings γ_A, γ_B and closings φ_A, φ_B verify:

$$\gamma_B \leq \gamma_A \iff \varphi_B \geq \varphi_A \iff B \sqsupseteq A \ .$$

In other words, the "greater" is the structuring element (for $\sqsupseteq$), the more active are the opening and closing.

The above suggests that the activity of openings and closings is governed by the size of the structuring elements that they use. We see below that it can be characterized in another way:

PROPOSITION 14.28 *Let ψ_1 and ψ_2 be either two openings, or two closings, on a poset L. Then the following four statements are equivalent:*

1. *ψ_2 is more active than ψ_1.*
2. *$Inv(\psi_2) \subseteq Inv(\psi_1)$.*
3. *$\psi_2\psi_1 = \psi_2$.*
4. *$\psi_1\psi_2 = \psi_2$.*

The second item indicates that the activity increases as the domain of invariance decreases. The last two suggest a notion of a filtering absorption order associated to activity: if ψ_2 is more active than ψ_1, then as a filter ψ_2 is more severe, so ψ_1 does not improve in any way upon the result of ψ_2, whether applied before of after it; thus ψ_2 *absorbs* ψ_1. We can now consider an ordered sequence of openings:

DEFINITION 14.29 *A* granulometry *(on a poset L) is a family of operators* ψ_r *(*$r \in R \subseteq \mathbb{R}^+$*) such that:*

1 $\forall r \in R$, ψ_r *is anti-extensive;*

2 $\forall r \in R$, ψ_r *is increasing;*

3 $\forall r, s \in R$, $\psi_r \psi_s = \psi_s \psi_r = \psi_{\max(r,s)}$.

Applying item 3 with $r = s$, ψ_r is idempotent, so it is an opening. In fact, ψ_r ($r \in R$) is a granulometry iff ψ_r is an opening for every r, and ψ_r decreases (becomes more active) as the parameter r increases: $\forall r, s \in R$, $r \geq s$ implies $\psi_r \leq \psi_s$, or equivalently $Inv(\psi_r) \subseteq Inv(\psi_s)$.

For binary images in a digital framework ($L = \mathcal{P}(\mathbb{Z}^n)$), we take $R = \{2, \ldots, r_{max}\}$ and ψ_r to be the opening by a structuring element B_r corresponding to size r (say, a $r \times r$-square). Then for a set $X \subseteq \mathbb{Z}^n$, it is interesting to measure the area (number of pixels) of $\gamma_r(X)$ for all r; this gives a decreasing function $R \to \mathbb{N}$, the *granulometry curve* of X, it displays the area of the objects according to the size of the opening. Positions where this curve decreases sharply indicate that there are substantial parts of X having the corresponding width. This is illustrated in Fig. 14.11.

One defines similarly an *anti-granulometry* by replacing, in item 1 of Definition 14.29, "anti-extensive" by "extensive". Then ψ_r ($r \in R$) is an anti-granulometry iff ψ_r is a closing for every r, and ψ_r increases (becomes more active) as the parameter r increases: $\forall r, s \in R$, $r \geq s$ implies $\psi_r \geq \psi_s$, or equivalently $Inv(\psi_r) \subseteq Inv(\psi_s)$. In a digital framework, we define the *anti-granulometry curve*, which gives an indication on the width of the holes of the set.

It is possible to combine a granulometry γ_r ($r \in R_1 \subseteq \mathbb{R}^+$) and an anti-granulometry φ_r ($r \in R_2 \subseteq \mathbb{R}^+$) into a two-sided sequence ψ_r, $r \in R = R_1 \cup \{0\} \cup (-R_2)$ by setting $\psi_r = \gamma_r$ for $r \in R_1$, $\psi_0 = \mathbf{id}$, and $\psi_{-r} = \varphi_r$ for $r \in R_2$. Then the axioms are: (1) ψ_r is anti-extensive for $r \geq 0$ but extensive for $r \leq 0$, (2) ψ_r is increasing, (3) $\psi_r \psi_s = \psi_s \psi_r = \psi_{m(r,s)}$ for r, s having the same sign, where $m(r, s) = \max(r, s)$ for $r, s \geq 0$, but $m(r, s) = \min(r, s)$ for $r, s \neq 0$ (we have no such identity for $r > 0$ and $s < 0$). We generalize then the granulometry curve into a function $R \to \mathbb{N}$ where the parts $r < 0$ and $r > 0$ deal with the sizes of holes and grains respectively.

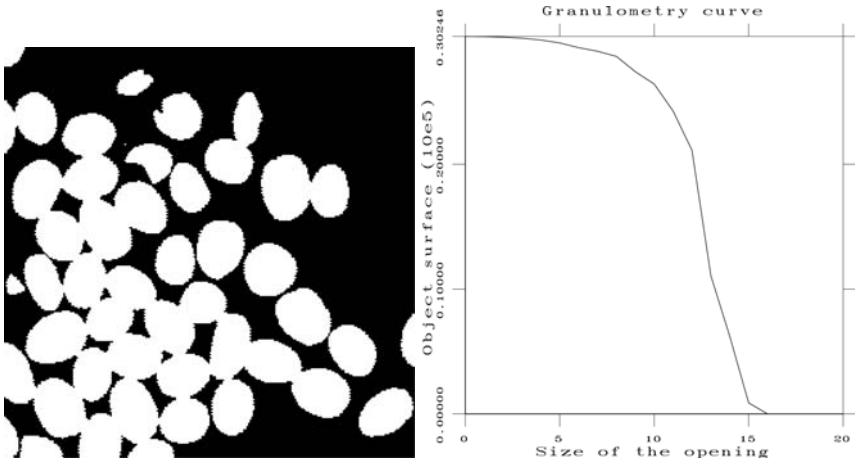


Figure 14.11. A binary image of coffee beans, and its granulometry curve, showing a sharp decrease between 10 and 15; this indicates that most beans have such a width.

3. Related approaches

As we said in the Introduction, there are other fields of research which are based on the same lattice-theoretical foundations as mathematical morphology. We present here three of them, which we think are relevant to the logic of spatial relations: formal concept analysis, rough sets and fuzzy sets. Roughness can be represented by using mathematical morphology operators to define upper and lower approximations in the framework of roughs sets (Sec. 3.2). We then show why mathematical morphology can be considered as a spatial reasoning tool (Sec. 3.3), with its two components: spatial knowledge representation and reasoning. As for the first one, we present in Sec. 3.4 an extension of mathematical morphology to fuzzy sets, which leads to an extended representation power coping with spatial imprecision. Modeling spatial relationships based on fuzzy morphology allows reasoning under imprecision and with structural spatial information, as shown in Sec. 3.5. This reasoning component will be further explored in Sec. 4.

3.1 Formal Concept Analysis

It is a lattice-based theory (Ganter and Wille, 1999) of relations between objects and features. It can be applied to spatial objects and spatial relations like visibility, enclosure, etc.

Let Ω be a set of *objects*, Π a set of *properties*, and $\sim$ a relation between Ω and Π, where $o \sim p$ means that object o has property p. The triple $(\Omega, \Pi, \sim)$ is

called a *context*. Following Theorem 14.12 we define the two maps

$$\begin{array}{lll} \omega: & \mathcal{P}(\Pi) \to \mathcal{P}(\Omega): & P \mapsto \{o \in \Omega \mid \forall p \in P,\ o \sim p\} \ , \\ \pi: & \mathcal{P}(\Omega) \to \mathcal{P}(\Pi): & O \mapsto \{p \in \Pi \mid \forall o \in O,\ o \sim p\} \ . \end{array}$$

Thus $\omega(P)$ is the set of objects sharing all properties in P, while $\pi(O)$ is the set of properties shared by all objects in O. A *concept* is a pair (O, P), where $O \in \mathcal{P}(\Omega)$ and $P \in \mathcal{P}(\Pi)$, such that $O = \omega(P)$ and $P = \pi(O)$; O is the *extent* of the concept and P is the *intent* of the concept.

The set of concepts ordered by inclusion of object sets, or equivalently by the inverse inclusion on property sets:

$$(O_1, P_1) \leq (O_2, P_2) \iff O_1 \subseteq O_2 \iff P_1 \supseteq P_2 \ .$$

It forms then a complete lattice with the following supremum and infimum operations:

$$\begin{array}{lll} \bigvee_{i \in I} (O_i, P_i) & = & \Big(\omega\pi\Big[\bigcup_{i \in I} O_i\Big], \bigcap_{i \in I} P_i\Big) \ , \\ \bigwedge_{i \in I} (O_i, P_i) & = & \Big(\bigcap_{i \in I} O_i, \pi\omega\Big[\bigcup_{i \in I} P_i\Big]\Big) \ . \end{array}$$

Here we used the complete lattice structure of the two Moore families of extents (possible O_i's) and of intents (possible P_i's), cf. Proposition 14.9 and Theorem 14.5.

A possible example of application of formal concept analysis is to consider two subsets S and V of the space $\mathbb{R}^n$ or $\mathbb{Z}^n$, where S represents an object to be visually examined, and a V is a set of viewpoints. We define a relation $\sim$ between the boundary ∂S of S and V, namely $s \sim v$ if s is visible from v, i.e., the open segment $]s, v[$ is disjoint from S. Then a concept is given by a pair (T, W), where $T \subseteq \partial S$ and $W \subseteq V$, such that T is the set of positions visible by all viewpoints in W, and W is the set of all viewpoints from which T is entirely visible.

3.2 Mathematical morphology and rough sets

Rough set theory has been introduced by Pawlak, 1982, as an extension of set theory, mainly in the domain of intelligent systems. The objective was to deal with incomplete information, leading to the idea of indistinguishability of objects in a set. It is therefore related to the concept of approximation, and of granularity of information (in the sense of Zadeh, 1979). This theory was applied successfully in several applications, e.g. information analysis, data analysis and data mining, knowledge discovery (for instance discovery of which features are relevant for data description), i.e., all applications for which a need arises for intelligent decision support. Let us mention in particular the works

of Lin, 1995, Lin and Liu, 1994, Yao, 1998 and Yao and Lin, 1996. There have also been studies towards a fuzzy approach to rough sets (Dubois and Prade, 1990), and on their relations with logic (Orlowska, 1993; Pawlak, 1987).

In this framework, a set X is approximated by two sets, called *upper* and *lower approximations*, and denoted by $\overline{A}(X)$ and $\underline{A}(X)$, such that $\underline{A}(X) \subseteq X \subseteq \overline{A}(X)$. It is interesting to investigate the algebraic properties of the two set operators $\overline{A}$ and $\underline{A}$: the first one is extensive, the second one is anti-extensive, and probably they should be increasing. But then, are they respectively a dilation and an erosion? In particular, do they constitute an adjunction (arising from a reflexive relation, cf. Proposition 14.21)? Or are they a closing and an opening? Or is $\overline{A}$ both a closing and a dilation, and $\underline{A}$ both an opening and an erosion (cf. Proposition 14.22)? Are they dual by complementation? Surprisingly, there have not been many studies on the relation between rough sets and MM; let us mention a few of them. On the one hand Polkowski, 1998 built a hit-or-miss (Fell) topology on rough sets, similar to the one used in MM for closed sets (Heijmans, 1994; Matheron, 1975; Ronse and Tajine, 2004; Serra, 1982). On the other hand Bloch, 2000b studied the algebraic properties of the upper and lower approximation operators, and established an analogy between them and the classical morphological operators on Euclidean or digital sets, namely dilation, erosion, opening and closing by a structuring element. Also Düntsch and Gediga, 2003 considered the algebraic aspects of rough sets, in particular their links with Galois connections.

Here we will investigate rough sets in light of the theory of adjunctions on sets; in some sense, this is a generalization of Bloch, 2000b. At the same time we will address topological aspects. But let us first recall the basic definitions of rough sets, in particular those based on a similarity relation.

In rough set theory (Pawlak, 1982), the two approximations $\overline{A}(X)$ and $\underline{A}(X)$ such that $\underline{A}(X) \subseteq X \subseteq \overline{A}(X)$ are defined from an equivalence relation. Let $\mathcal{U}$ denote the universe of discourse (X being a subset of $\mathcal{U}$). We consider *attributes* which are functions defined on $\mathcal{U}$, and write A for the set of attributes. To each $x \in \mathcal{U}$ we associate an *information vector* $Inf(x)$, which is the set of attributes associated to x. We define an equivalence relation R_A (with respect to the set A of attributes on $\mathcal{U}$) by the equality of the information vector:

$$x \; R_A \; y \iff Inf(x) = Inf(y) \;\; .$$

Assuming that each element of $\mathcal{U}$ is known only through its attributes, $x \; R_A \; y$ means that x and y are undistinguishable on the basis of available information. The pair $(\mathcal{U}, R_A)$ is called an approximation space. For $x \in \mathcal{U}$, let $[x]_A$ denote the equivalence class of x under R_A:

$$[x]_A = \{y \in \mathcal{U} \mid x \; R_A \; y\} \;\; .$$

Then upper and lower approximations of a subset X of $\mathcal{U}$ are defined as:

$$\begin{array}{rcl} \overline{A}(X) & = & \{x \in \mathcal{U} \mid [x]_A \cap X \neq \emptyset\} \ , \\ \underline{A}(X) & = & \{x \in \mathcal{U} \mid [x]_A \subseteq X\} \ . \end{array} \tag{14.29}$$

The lower approximation of X contains all points of $\mathcal{U}$ that are distinguishable from every elements of X^c, while its upper approximation contains all points of $\mathcal{U}$ that are undistinguishable from some element of X. We call a *rough set* a pair $(\underline{A}(X), \overline{A}(X))$.

Let us refer to the terminology used for dilations and erosions on sets: if R_A stands for ρ, then by Eqs. (14.24,14.25,14.26) we have $[x]_A = N(x)$, $\underline{A} = \varepsilon_N = \varepsilon_{R_A}$ and $\overline{A} = \delta_{\widetilde{N}} = \delta_{R_A^{-1}}$. Clearly $\overline{A}$ and $\underline{A}$ are dual under complementation: $\overline{A}(X) = \left[\underline{A}(X^c)\right]^c$. The fact that R_A is symmetrical ($R_A = R_A^{-1}$) means that $N = \widetilde{N}$, so $(\underline{A}, \overline{A})$ forms an adjunction. Now as R_A is reflexive and transitive, by Proposition 14.22, the erosion $\underline{A}$ is also an opening, while the dilation $\overline{A}$ is also a closing, with $\overline{A}\underline{A} = \underline{A}$ and $\underline{A}\overline{A} = \overline{A}$. In particular, we have

$$\forall X \in \mathcal{P}(\mathcal{U}), \qquad \underline{A}(X) \subseteq X \subseteq \overline{A}(X) \ .$$

By Proposition 14.22 again, $\varepsilon_N(\mathcal{P}(\mathcal{U})) = \delta_N(\mathcal{P}(\mathcal{U}))$, i.e., the families of lower and upper approximations coincide:

$$\{\underline{A}(X) \mid X \in \mathcal{P}(\mathcal{U})\} = \{\overline{A}(X) \mid X \in \mathcal{P}(\mathcal{U})\} \ ;$$

in fact this family consists of all sets which are unions of equivalence classes $[x]_A$. With the topological interpretation given after Proposition 14.22, it constitutes an Alexandroff topology on $\mathcal{U}$, where open sets coincide with closed ones, the upper approximation $\overline{A}(X)$ is the closure and the star (least open superset) of X, while the lower approximation $\underline{A}(X)$ is the interior and the greatest closed subset of X.

This definition can be extended to any relation R, leading to the notion of generalized approximate space (see e.g. Yao, 1998). Simply we take an arbitrary relation R instead of the equivalence R_A, and the set

$$r(x) = \{y \in \mathcal{U} \mid x \ R \ y\}$$

instead of the equivalence class $[x]_A$; here $r(x)$ corresponds to the set $N(x)$ according to Equation (14.25), with R standing for ρ. Then Equation (14.29) becomes

$$\begin{array}{rcl} \overline{R}(X) & = & \{x \in \mathcal{U} \mid r(x) \cap X \neq \emptyset\} \ , \\ \underline{R}(X) & = & \{x \in \mathcal{U} \mid r(x) \subseteq X\} \ . \end{array} \tag{14.30}$$

In our terminology, $\underline{R} = \varepsilon_r = \varepsilon_R$ and $\overline{R} = \delta_{\widetilde{r}} = \delta_{R^{-1}}$. Now $\overline{R}$ and $\underline{R}$ are still dual under complementation.

Equivalently, we can define $\underline{R}$ and $\overline{R}$ as two set operators which are dual under complementation ($\overline{R}(X) = \left[\underline{R}(X^c)\right]^c$), and such that $\underline{R}$ is an erosion (or equivalently: $\overline{R}$ is a dilation). This is in accordance with the operator-oriented view of rough sets (Lin and Liu, 1994; Yao, 1998).

If R is an equivalence relation, we get Pawlak's definition, Equation (14.29); indeed, we can define $Inf(x) = [x]_R$, the equivalence class of x under R. Let us consider weaker conditions on R. We require that $\underline{R}(X) \subseteq \overline{R}(X)$ for all $X \in \mathcal{P}(\mathcal{U})$, which means that $\varepsilon_r \leq \delta_{\widetilde{r}}$; according to Proposition 14.20, this is verified iff

$$\forall x \in \mathcal{U}, \qquad r(x) \neq \emptyset.$$

Usually, one requires that $\underline{R}$ is anti-extensive and $\overline{R}$ is extensive, that is,

$$\forall X \in \mathcal{P}(\mathcal{U}), \qquad \underline{R}(X) \subseteq X \subseteq \overline{R}(X);$$

by Proposition 14.21, this is verified iff R is reflexive.

By Proposition 14.21, R is symmetrical iff $\underline{R}\overline{R}$ is extensive, iff $\overline{R}\underline{R}$ is anti-extensive. Then $r = \widetilde{r}$, and $(\underline{R}, \overline{R})$ is an adjunction. If R is reflexive and symmetrical, we call it a *tolerance relation*.

By Proposition 14.22, R is reflexive and transitive iff $\underline{R}$ is both an erosion and an opening, iff $\overline{R}$ is both a dilation and a closing. Here we have an Alexandroff topology on $\mathcal{U}$, where for every $X \in \mathcal{P}(\mathcal{U})$, $\underline{R}(X)$ is the topological interior of X and $\overline{R}(X)$ is the topological closure of X. The family of open sets, $\{\underline{R}(X) \mid X \in \mathcal{P}(\mathcal{U})\}$, and the family of closed sets,$\{\overline{R}(X) \mid X \in \mathcal{P}(\mathcal{U})\}$, do not coincide, unless R is also symmetrical (i.e., an equivalence relation).

When R is reflexive but not transitive, the upper and lower approximations do generally not correspond to a topology. However they correspond to the closure and interior in a *pre-topology*, that is: $\overline{R}$ and $\underline{R}$ are dual under complementation, they are both increasing, $\overline{R}$ is extensive while $\underline{R}$ is anti-extensive, $\overline{R}(\emptyset) = \emptyset$ and $\underline{R}(\mathcal{U}) = \mathcal{U}$. Let us mention the use by Emptoz, 1983 of pre-topology for the description of spatial objects. This may be of interest for pattern recognition purposes, since a non-idempotent closure allows to aggregate patterns using iterated closure operations.

In Yao, 1998, various properties are given for the operators $\underline{R}$ and $\overline{R}$, which may or may not be satisfied, according to the properties of the relation R. In Bloch, 2000b, a parallel is made between these properties and those of dilations, erosions, openings and closings. In fact, *all* these properties follow from the ones given in Sec. 2.3 for dilations and erosions on sets, those of openings and closings in Sec. 2.2, and from Equation (14.15).

Using the operator-oriented point of view (Lin and Liu, 1994; Yao, 1998), one could also define the lower and upper approximation as an opening and a closing. However, such operators cannot be defined in terms of a relation or a neighbourhood function, as was the case with dilations and erosions. Openings

and closings are characterized by their invariance domain: for an opening, it is a dual Moore family, i.e., a family closed under arbitrary unions and containing $\emptyset$; for a closing it is a Moore family, i.e., a family closed under arbitrary intersections and containing $\mathcal{U}$. Let L be an opening and let $\mathcal{S}$ be a family of nonvoid parts of $\mathcal{U}$ such that $Inv(L)$ is the set of all arbitrary unions of elements of $\mathcal{S}$ (including the empty union $\emptyset$); in fact L is the opening $\gamma_{\mathcal{S}}$ defined after Eqs. (14.18,14.19), as for every $X \in \mathcal{P}(\mathcal{U})$ we have

$$L(X) = \bigvee\{A \in \mathcal{S} \mid A \subseteq X\}.$$

For example, if $\mathcal{U} = E$ and L is the opening by a structuring element B, then $\mathcal{S}$ is the set of translates B_p ($p \in E$) of B. Now let H be the closing which is the dual of L by complementation; then $Inv(H) = \{A^c \mid A \in Inv(L)\}$, and $Inv(H)$ is the set of arbitrary intersections of A^c for $A \in \mathcal{S}$. For $X \in \mathcal{P}(\mathcal{U})$ we have

$$H(X) = \left(\bigvee\{A \in \mathcal{S} \mid A \subseteq X^c\}\right)^c.$$

We can interpret the elements of $\mathcal{S}$ as "blocks", and a point x can be included in the lower approximation $L(X)$ only through its membership of a "block" included in X, while a point x can be excluded from the upper approximation $H(X)$ only through its membership of a "block" excluded from X. However these blocks do not make a partition, as with Pawlak's definition, Equation (14.29).

An interesting particular case is when L and H are the interior and closure operators in a topological space. One speaks then of a *topological approximation space*. As explained after Proposition 14.22, this is weaker than requiring H to be both a dilation and a closing, and dually L to be both an erosion and an opening.

Other operators could be used for lower and upper approximations. In Serra, 1988, it is shown that every increasing operator on a complete lattice, which fixes the greatest element, is a supremum of erosions. Thus an increasing operator ψ on $\mathcal{P}(\mathcal{U})$ such that $\psi(\mathcal{U}) = \mathcal{U}$, is a union of erosions; in particular ψ is anti-extensive iff these erosions are anti-extensive. Consider the following definition of lower and upper approximations (Lin, 1995). Suppose that to each $x \in \mathcal{U}$ we associate a family $\mathcal{N}(x)$ of parts of $\mathcal{U}$ which are "neighbourhoods" of x; $\mathcal{N}(x)$ is called a *neighbourhood system of* x. Now we define the upper and lower approximations $\overline{N}$ and $\underline{N}$ as follows:

$$\begin{aligned} \overline{N}(X) &= \{x \in \mathcal{U} \mid \forall A \in \mathcal{N}(x), A \cap X \neq \emptyset\}, \\ \underline{N}(X) &= \{x \in \mathcal{U} \mid \exists A \in \mathcal{N}(x), A \subseteq X\}. \end{aligned} \tag{14.31}$$

Clearly $\overline{N}$ is the dual by complementation of $\underline{N}$, so let us analyse the latter. We have

$$\underline{N}(\mathcal{U}) = \{x \in \mathcal{U} \mid \mathcal{N}(x) \neq \emptyset\},$$

so in order to have $\underline{N}(\mathcal{U}) = \mathcal{U}$, we suppose that $\mathcal{N}(x) \neq \emptyset$ for all $x \in \mathcal{U}$. Then $\underline{N}$ is a union of erosions, and we can describe them precisely. For every $x \in \mathcal{U}$ and for every $A \in \mathcal{N}(x)$, let $N[x, A] : \mathcal{U} \to \mathcal{P}(\mathcal{U})$ be defined by $N[x, A](x) = A$ and $N[x, A](y) = \mathcal{U}$ for $y \in \mathcal{U} \setminus \{x\}$. Then $\varepsilon_{N[x,A]}$ verifies for every $X \in \mathcal{P}(\mathcal{U})$:

$$\varepsilon_{N[x,A]}(X) = \begin{cases} \mathcal{U} & \text{if } X = \mathcal{U}, \\ \{x\} & \text{if } A \subseteq X \neq \mathcal{U}, \\ \emptyset & \text{otherwise.} \end{cases}$$

We obtain then

$$\underline{N}(X) = \bigcup_{x \in \mathcal{U}} \bigcup_{A \in \mathcal{N}(x)} \varepsilon_{N[x,A]}(X).$$

Dually, we get

$$\overline{N}(X) = \bigcap_{x \in \mathcal{U}} \bigcap_{A \in \mathcal{N}(x)} \delta_{\widetilde{N[x,A]}}(X).$$

This view is particularly interesting for shape recognition, since in morphological recognition, an object has often to be tested or matched with a set of patterns, like directional structuring elements. Thus we apply the union of the erosions by all those patterns, which is a particular case of the above operator.

In Bloch, 2000b, some other extensions are presented, using as lower and upper approximations morphological thinning and thickening (Serra, 1982). There is also an extension to rough functions, using the grey-level morphological operations described in Sec. 1.2.

To conclude, let us remark that the general theory of adjunctions, dilations, erosions, openings and closings on sets provides a good formal framework for expressing the notion of coarseness underlying rough sets. It allows to characterize precisely their algebraic properties and their relations with topology. In Sec. 3.4 and 4, we will examine the relationship of rough sets with fuzzy sets and modal logic, especially in the morphological framework.

3.3 Mathematical morphology and spatial reasoning

Spatial reasoning has been largely developed in artificial intelligence, in particular using qualitative representations based on logical formalisms. In image interpretation and computer vision it is much less developed and is mainly based on quantitative representations. On the contrary, mathematical morphology is widely used in these domains. A typical example concerns model-based structure recognition in images, where the model represents spatial entitites and relationships between them. Based on this example, spatial reasoning can be defined as the domain of spatial knowledge representation, in particular spatial relations between spatial entities, and of reasoning on these entities and

relations. This definition exhibits two main components: spatial knowledge representation and reasoning. In particular spatial relationships constitute an important part of the knowledge we have to handle and imprecision is often attached to it. The reasoning component includes fusion of heterogeneous spatial knowledge, decision making, inference, recognition. Two types of questions are raised when dealing with spatial relationships:

1 given two objects (possibly fuzzy), assess the degree to which a relation is satisfied;

2 given one reference object, define the area of space in which a relation to this reference is satisfied (to some degree).

In order to answer these questions and address both representation and reasoning issues, different frameworks and their combination can be used. Fuzzy set theory has powerful features to represent imprecision at different levels, to combine heterogeneous information and to make decision (Dubois and Prade, 1985; Dubois et al., 1999). Formal logics and the attached reasoning and inference power are widely used too, usually in a qualitative context. But mathematical morphology, which is an algebraic theory that has extensions to fuzzy sets and to logical formulas, is a very promising tool, since it can elegantly unify the representation of several types of relationships (Bloch, 2003b). The association of different frameworks for spatial reasoning allows us to match two requirements such as axiomatization, expressiveness and completeness with respect to the types of spatial information we want to represent (Aiello, 2002). Complexity issues are not addressed here, but it should be noted that efficient algorithms exist for digital morphology.

Mathematical morphology provides tools for spatial reasoning at several levels. It provides tools for representing objects or object features (see e.g. Sec. 1.1 and 2). For instance skeletons provide compact and expressive representations of shapes; morphological tools for shape decomposition lead to structured representations, such as graphs for instance; spatial imprecision can be represented by a pair of dilation and erosion; tools for selecting objects or parts of objects having specific properties can be derived from morphological operators such as hit-or-miss transformations for instance, etc. These aspects, quite traditional in mathematical morphology, are not detailed here, and we will concentrate rather on tools for representing spatial relations. The notion of structuring element captures the local spatial context and leads to analysis of a scene using operators involving the neighbourhood of each point. At a more global level, several spatial relations between spatial entities can be expressed as morphological operations, in particular using dilations. Let us provide a few examples, of metric and topological relationships.

Several distances between objects can be expressed in terms of dilation. The minimum or nearest point distance between two sets X and Y is defined (in the

discrete finite case) as:

$$\begin{aligned} d_N(X,Y) &= \min_{(x,y)\in X\times Y} d_E(x,y) \\ &= \min_{x\in X} d_E(x,Y) = \min_{y\in Y} d_E(y,X), \end{aligned} \tag{14.32}$$

where d_E denotes the Euclidean distance in $\mathcal{S}$ (note that this function is improperly named distance since it is not separable and does not satisfy the triangular inequality). This has an equivalent morphological expression:

$$\begin{aligned} d_N(X,Y) &= \inf\{n \in \mathbb{N}, X \cap \delta_n(Y) \neq \emptyset\} \\ &= \inf\{n \in \mathbb{N}, Y \cap \delta_n(X) \neq \emptyset\}. \end{aligned} \tag{14.33}$$

Another morphological expression is, for $n > 0$:

$$d_N(X,Y) = n \Leftrightarrow \delta_n(X) \cap Y \neq \emptyset \text{ and } \delta_{(n-1)}(X) \cap Y = \emptyset \tag{14.34}$$

or equivalently the symmetrical expression. For $n = 0$ we have:

$$d_N(X,Y) = 0 \Leftrightarrow X \cap Y \neq \emptyset. \tag{14.35}$$

The Hausdorff distance is defined as:

$$H_d(X,Y) = \max[\sup_{x\in X} d_E(x,Y), \sup_{y\in Y} d_E(y,X)]. \tag{14.36}$$

Similarly as for the nearest point distance, this distance can be expressed in morphological terms as:

$$H_d(X,Y) = \inf\{n, X \subseteq \delta_n(Y) \text{ and } Y \subseteq \delta_n(X)\}. \tag{14.37}$$

Alternatively, we can write:

$$H_d(X,Y) = 0 \Leftrightarrow X = Y, \tag{14.38}$$

and for $n > 0$:

$$\begin{aligned} &H_d(X,Y) = n \Leftrightarrow X \subseteq \delta_n(Y) \text{ and } Y \subseteq \delta_n(X) \\ &\quad \text{and } \left(X \not\subseteq \delta_{(n-1)}(Y) \text{ or } Y \not\subseteq \delta_{(n-1)}(X)\right). \end{aligned} \tag{14.39}$$

From these representations, several types of knowledge about distance can be expressed. For instance, Fig. 14.12 shows a spatial representation of "B is at a distance between n_1 and n_2 from A", i.e. B should be in the dilation of radius n_2 of A but not in the dilation of radius n_1 of A.

Another example is adjacency. Here, we restrict ourselves to the digital case, and use discrete topology as derived from digital connectivity for defining

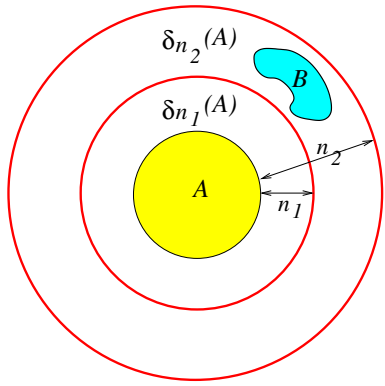


Figure 14.12. Illustration of a distance relation expressed by an interval.

adjacency between two regions X and Y, subsets of the digital space (see Ch. 12 of this book for further details about digital topology). Let us consider an n-dimensional digital space (typically $\mathbb{Z}^n$), and any discrete connectivity defined on this space, denoted c-connectivity (for instance, for $n = 3$, we may consider 6-, 18- or 26-connectivity on a cubic grid). Let $n_c(x, y)$ be the Boolean variable stating that x and y are neighbours in the sense of the discrete c-connectivity. Let B_c be the set of c-neighbours of the origin. For any two subsets X and Y in $\mathbb{Z}^n$, X and Y are adjacent according to the c-connectivity if: $X \cap Y = \emptyset \;\; and \;\; \exists x \in X, \exists y \in Y : n_c(x, y)$.

This definition can also be expressed equivalently in terms of morphological dilation, as:

$$X \cap Y = \emptyset \;\text{ and }\; \delta_{B_c}(X) \cap Y \neq \emptyset, \tag{14.40}$$

where $\delta_{B_c}(X)$ denotes the dilation of X by the structuring element B_c.

Another topological relation, often used in the context of mereotopology for instance, is tangential proper part. Again this can be expressed in morphological terms, as illustrated in Fig. 14.13.

These expressions extend to different frameworks, including fuzzy set theory and formal logics, thus benefiting from the reasoning power of these frameworks.

3.4 Fuzzy mathematical morphology

Dealing with spatial imprecision can be adequately addressed by defining objects or regions of space as fuzzy objects. We denote by $\mathcal{S}$ the spatial domain, typically $\mathbb{Z}^2$ or $\mathbb{Z}^3$ for digital 2D or 3D images, or, in the continuous case, $\mathbb{R}^2$ or $\mathbb{R}^3$. A fuzzy object is a fuzzy set defined on $\mathcal{S}$, i.e. a spatial fuzzy set. Its

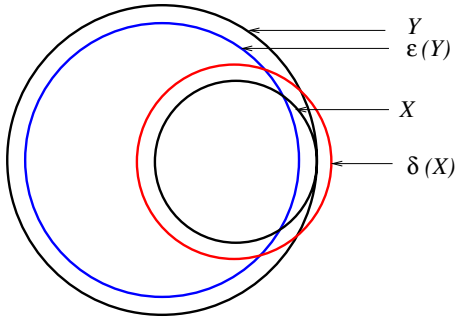


Figure 14.13. Illustration of tangential part relationship, and its expression in terms of dilation and erosion: X is included in Y while its dilation is not (equivalently, X is not included in the erosion of Y).

membership function μ represents the imprecision in the spatial extent of the object. For any point x of $\mathcal{S}$, $\mu(x)$ is the degree to which x belongs to the fuzzy object. We denote by $\mathcal{F}$ the set of fuzzy sets on $\mathcal{S}$.

Several definitions of fuzzy mathematical morphology have been proposed since a few years. Some of them just consider grey level as membership functions (Goetcherian, 1980; Giardina and Sinha, 1989; Laplante and Giardina, 1991; di Gesu, 1988; di Gesu et al., 1993; Nakatsuyama, 1993), or use binary structuring elements (Rosenfeld, 1984). Here we restrict the presentation to really fuzzy approaches, where fuzzy sets have to be transformed according to fuzzy structuring elements. Initial developments can be found in the definition of fuzzy Minkowski addition (Dubois and Prade, 1983; Kaufmann and Gupta, 1988). Then this problem has been addressed by several authors independently (Bloch, 1993; Bloch and Maître, 1995; Sinha and Dougherty, 1992; de Baets and Kerre, 1993; de Baets, 1995; Bandemer and Näther, 1992; Popov, 1995; Sinha et al., 1997; Nachtegael and Kerre, 2000; Maragos et al., 2001; Deng and Heijmans, 2002).

Attention will be paid here only to the 4 basic operations of mathematical morphology (erosion, dilation, opening, closing), but it should be clear for the reader that for every definition, a complete set of morphological operations could be derived. Extensions of mathematical morphology have been proposed for instance for defining more complex operations (like filtering) (Bloch and Maître, 1995; Sinha et al., 1997), and geodesic operations (Bloch, 2000a).

Although a fuzzy set is defined through its membership function, functional approaches are not appropriate. For instance, the classical dilation of a function taking values in $[0, 1]$ by a functional structuring element taking values in $[0, 1]$ generally provides a function with values in $[0, 2]$ which has no direct interpretation in terms of fuzzy sets. Therefore a set theoretical approach is preferred,

where set operations are converted into their fuzzy equivalents, thus preserving the compatibility with classical morphology in case the fuzzy sets reduce to crisp ones.

General methods for extending an operation to fuzzy sets. Common and generic methods that can be used for defining a fuzzy operator or fuzzy relationship from the corresponding binary ones can be categorized in three main classes. The first type relies on the "extension principle", as introduced by Zadeh (Zadeh, 1975; Dubois and Prade, 1980). The second class relies on computation on α-cuts (e.g. Dubois and Jaulent, 1987; Krishnapuram et al., 1993; Bloch and Maître, 1995). These two classes of definitions explicitly involve the operations or relations on crisp sets. The third class of methods consists in providing directly fuzzy definitions of the operations or of the relationships, by substituting all crisp expressions by their fuzzy equivalents. This type of translation is used in the following.

This translation is generally done term by term. For instance, intersection is replaced by a t-norm, union by a t-conorm, sets by fuzzy set membership functions, etc. A triangular norm (or t-norm) is a function from $[0,1] \times [0,1]$ into $[0,1]$ which is commutative, associative, increasing, and for which 1 is unit element and 0 is null element (Menger, 1942; Schweizer and Sklar, 1963). Examples of t-norms are min, product, etc. (Dubois and Prade, 1980). A t-conorm is its dual with respect to complementation. This type of translation is particularly straightforward if the binary relationship can be expressed in set theoretical and logical terms.

Let us take a simple example to illustrate this method. A fuzzy set μ is said to be included in another fuzzy set ν if $\forall x \in \mathcal{S},\ \mu(x) \leq \nu(x)$. This is a crisp definition of inclusion of fuzzy sets. We may also consider that if two sets are imprecisely defined, their inclusion relationship may be imprecise too. Therefore inclusion of fuzzy sets becomes a matter of degree. This degree of inclusion can be obtained using the translation principle. In the crisp case, the set equation expressing inclusion of a set X in a set Y can be written as follows:

$$X \subseteq Y \Leftrightarrow X^c \cup Y = \mathcal{S} \Leftrightarrow \forall x \in \mathcal{S},\ x \in X^c \cup Y. \tag{14.41}$$

In the fuzzy case, X and Y become fuzzy sets having membership functions μ and ν and we have the following correspondences:

$$\forall x \in \mathcal{S} \quad \leftrightarrow \quad \inf_{x \in \mathcal{S}}, \tag{14.42}$$

$$x \in X^c \quad \leftrightarrow \quad c[\mu(x)], \tag{14.43}$$

$$x \in Y \quad \leftrightarrow \quad \nu(x), \tag{14.44}$$

$$X^c \cup Y \quad \leftrightarrow \quad T[c(\mu), \nu]. \tag{14.45}$$

Finally, the degree of inclusion of μ in ν is defined as:

$$\mathcal{I}(\mu,\nu) = \inf_{x\in\mathcal{S}} T[c(\mu(x)),\nu(x)], \tag{14.46}$$

where T is a t-conorm and c a fuzzy complementation.

Fuzzy morphology by formal translation based on t-norms and t-conorms. The first attempts to build fuzzy mathematical morphology based on this translation principle were developed in Bloch, 1993 and Bloch and Maître, 1995, and coincide with the definitions independently developed in Bandemer and Näther, 1992. An important property that was put to the fore in this approach is the duality between erosion and dilation.

From the following set equivalence (where $\varepsilon_B(X)$ denotes the erosion of the set X by B): $x \in \varepsilon_B(X) \Leftrightarrow B_x \subseteq X$, a natural way to define the erosion of a fuzzy set μ by a fuzzy structuring element ν is to use the degree of inclusion defined above:

$$\forall x \in \mathcal{S},\ \varepsilon_\nu(\mu)(x) = \inf_{y\in\mathcal{S}} T[c(\nu(y-x)),\mu(y)]. \tag{14.47}$$

By duality with respect to the complementation c, fuzzy dilation is then defined by:

$$\forall x \in \mathcal{S},\ \delta_\nu(\mu)(x) = \sup_{y\in\mathcal{S}} t[\nu(x-y),\mu(y)], \tag{14.48}$$

where t is the t-norm associated to the t-conorm T with respect to the complementation c . This definition of dilation corresponds to the following set equivalence:

$$x \in \delta_B(x) \Leftrightarrow \check{B}_x \cap X \neq \emptyset \Leftrightarrow \exists y \in \mathcal{S},\ y \in \check{B}_x \cap X. \tag{14.49}$$

Here, intersection $\cap$ has been translated in terms of a t-norm t and the existential symbol by a supremum.

This form of fuzzy dilation and fuzzy erosion are very general, and several definitions found in the literature appear as particular cases (such as Bandemer and Näther, 1992; de Baets and Kerre, 1993; Sinha and Dougherty, 1992; Rosenfeld, 1984; Kaufmann and Gupta, 1988; Goetcherian, 1980).

Finally, fuzzy opening (respectively fuzzy closing) is simply defined as the combination of a fuzzy erosion followed by a fuzzy dilation (respectively a fuzzy dilation followed by a fuzzy erosion), by using dual t-norms and t-conorms.

Weak t-norms and t-conorms are weaker forms of t-norms and t-conorms: they are not associative and do not admit 1 (respectively 0) as unit element, in general. If we replace t-norms and t-conorms by these weaker forms in the previous construction, then Eqs. (14.47,14.48) appear as a generalization of the definitions proposed in Sinha and Dougherty, 1993. But they lead to weaker properties, and are therefore somewhat less interesting from a morphological point of view.

Properties of basic fuzzy morphological operations. The detail of properties for various definitions can be found in Bloch and Maître, 1995. We summarize here the main properties when using t-norms and t-conorms:

- duality of erosion and dilation (respectively opening and closing) with respect to the complementation c;
- compatibility with classical morphology if the structuring element is binary;
- translation-invariance (see Sec. 2);
- local knowledge property;
- continuity if the t-norm is continuous (which is most often the case);
- increasingness of all operations with respect to inclusion;
- extensivity of dilation and anti-extensivity of erosion iff $\nu(0) = 1$ (this corresponds to the condition that the origin should belong to the structuring element in the crisp case);
- extensivity of closing, anti-extensivity of opening and idempotence of these two operations iff $t[b, T(c(b), a)] \leq a$, which is satisfied for Lukasiewicz t-norm ($t(a, b) = \max(0, a+b-1)$) and t-conorm ($T(a, b) = \min(1, a + b)$);
- commutation of dilation with union (and of erosion with intersection);
- iteration property of dilation ($\delta_r \delta_s = \delta_{r+s}$) and of erosion.

Fuzzy morphology using adjunction and residual implications. A second type of approach is based on the notion of adjunction and fuzzy implications. Here the algebraic framework is the main guideline, which contrasts with the previous approach where duality was imposed in first place.

Fuzzy implication is often defined as (Dubois and Prade, 1991):

$$Imp(a, b) = T[c(a), b]. \tag{14.50}$$

Fuzzy inclusion is related to implication by the following equation:

$$\mathcal{I}(\nu, \mu) = \inf_{x \in \mathcal{S}} Imp[\nu(x), \mu(x)], \tag{14.51}$$

which allows to relate directly fuzzy erosion to fuzzy implication, leading to the general definition using t-conorm, and by duality also fuzzy dilation.

This suggests another way to define fuzzy erosion (and dilation), by using other forms of fuzzy implication. One interesting approach is to use residual implications:

$$Imp(a,b) = \sup\{\varepsilon \in [0,1], t(a,\varepsilon) \leq b\}. \tag{14.52}$$

This provides the following expression for the degree of inclusion:

$$\mathcal{I}(\nu,\mu) = \inf_{x\in\mathcal{S}} \sup\{\varepsilon \in [0,1], t(\nu(x),\varepsilon) \leq \mu(x)\}. \tag{14.53}$$

This definition coincides with the previous one for particular forms of t, typically Lukasiewicz t-norm.

The derivation of fuzzy morphological operators from residual implication has been proposed in de Baets, 1995, and then developed e.g. in de Baets, 1997 and Nachtegael and Kerre, 2000. One of its main advantages is that it leads to idempotent fuzzy closing and opening. This approach was formalized from the algebraic point of view of adjunction, as developed in Deng and Heijmans, 2002. This approach has then been used by other authors, such as in Maragos et al., 2001 and Maragos, 2005. This leads to general algebraic fuzzy erosion and dilation. Let us detail this approach. A fuzzy implication I is a mapping from $[0,1]\times[0,1]$ into $[0,1]$ which is decreasing in the first argument, increasing in the second one and satisfies $I(0,0) = I(0,1) = I(1,1) = 1$ and $I(1,0) = 0$. A fuzzy conjunction is a mapping from $[0,1] \times [0,1]$ into $[0,1]$ which is increasing in both arguments and satisfies $C(0,0) = C(1,0) = C(0,1) = 0$ and $I(1,1) = 1$. If C is also associative and commutative, it is a t-norm. A pair of operators (I,C) are said adjoint if:

$$C(a,b) \leq c \Leftrightarrow b \leq I(a,c). \tag{14.54}$$

The adjoint of a conjunction is a residual implication.

Fuzzy dilation and erosion are then defined as:

$$\delta_\nu(\mu)(x) = \sup_y C(\nu(x-y),\mu(y)), \tag{14.55}$$

$$\varepsilon_\nu(\mu)(x) = \inf_y I(\nu(y-x),\mu(y)). \tag{14.56}$$

Note that (I,C) is an adjunction if and only if $(\varepsilon_\nu,\delta_\nu)$ is an adjunction.

Opening and closing derived from these operations by combination have all required properties, whatever the choice of C and I. Some properties of dilation and erosion, such as iterativity, require C and I to be associative and commutative.

If C is a t-norm, then the dilation is exactly the same as the one obtained in the first approach. To understand the relation between both approaches for

erosion, we define $\hat{I}(a,b) = I(c(a),b)$ where c is a fuzzy complementation. In the following, we simply take $c(a) = 1-a$ which is the most usual complementation. Then $\hat{I}$ is increasing in both arguments, and if I is further assumed to be commutative and associative, $\hat{I}$ is a t-conorm. Equation (14.56) can be rewritten as: $\varepsilon_\nu(\mu)(x) = \inf_y \hat{I}(1-\nu(y-x),\mu(y))$, which corresponds to the fuzzy erosion of the first approach. The adjunction property can also be written as $C(a,b) \leq c \Leftrightarrow b \leq \hat{I}(1-a,c)$. However, pairs of dual t-norms and t-conorms are not identical to pairs of adjoint operators. Let us take a few examples. For $C = \min$, its adjoint is $I(a,b) = b$ if $b < a$, and 1 otherwise. But the derived $\hat{I}$ is the dual of the conjunction defined as $C(a,b) = 0$ if $b \leq 1-a$ and b otherwise. Conversely, the adjoint of this conjunction is $I(a,b) = \max(1-a,b)$, the dual of which is the minimum conjunction. Lukasiewicz operators are both adjoint and dual, which explains the exact correspondence between both approaches for these operators. Moreover, it can be proved that the condition for t-norms and t-conorms leading to idempotent opening and closing (i.e. $t(b,T(c(b),a) \leq a)$ is equivalent to the adjunction property between C and I for $t = C$ and $T = \hat{I}$. This new result completes the link between both approaches.

Fuzzy rough sets. We can now extend the links between rough sets and morphological operators derived in Sec. 3.2 to fuzzy rough sets. Using fuzzy mathematical morphology operators leads to fuzzy rough sets that have exactly the same properties as crisp rough sets, at least for particular t-norms and t-conorms (Bloch, 2000b). It turns out that these definitions using fuzzy erosion and dilation are generalizations of the ones proposed in Dubois and Prade, 1990, for $t = \min$ and $T = \max$ in a completely different context, using a fuzzy relation μ_R. The equivalence is obtained as in the crisp case by setting:

$$\mu_R(x,y) = \nu(y-x). \tag{14.57}$$

The interpretation is similar as in the crisp case: the degree of relation between x and y is equal to the degree to which $y-x$ belongs to the structuring element, i.e. to the degree to which y belongs to the structuring element translated at x. This approach has also been used in Nachtegael et al., 2000.

This extension brings together three different aspects of the information: rough sets represent coarseness, fuzzy sets represent vagueness and mathematical morphology brings a geometrical, topological and morphological aspect.

3.5 Spatial relationships and spatial reasoning from fuzzy mathematical morphology

Fuzzy distances derived from fuzzy dilation. The importance of distances in spatial reasoning is well established. Their extensions to fuzzy sets can be useful for dealing with imprecision and reasoning with semi-qualitative

(or semi-quantitative) information. Several definitions can be found in the literature for distances between fuzzy sets (which is the main addressed problem). They can be roughly divided into two classes (see Bloch, 2003a for a review): distances that take only membership functions into account and that compare them pointwise, and distances that additionally include spatial distances. The definitions which combine spatial distance and fuzzy membership comparison allow for a more general analysis of structures in space, for applications where topological and spatial arrangement of the structures of interest is important (such as spatial reasoning).

Morphological dilations are a convenient tool to define distances in the second class (Bloch, 1999b). The relations described in Sec. 3.3 express distances in set theoretical terms, and are therefore easier to translate with nice properties than usual analytical expressions. We detail the examples of nearest point distance and Hausdorff distance.

Fuzzy nearest point distance. By translating Equation (14.33), we define a distance distribution (Rosenfeld, 1985) $\Delta_N(\mu, \mu')(n)$ that expresses the degree to which the distance between μ and μ' is less than n by:

$$\Delta_N(\mu, \mu')(n) = f[\sup_{x \in \mathcal{S}} t[\mu(x), \delta_n(\mu')(x)], \sup_{x \in \mathcal{S}} t[\mu'(x), \delta_n(\mu)(x)]], \tag{14.58}$$

where δ_n is a fuzzy dilation of radius n ($\delta_n = (\delta_1)^n$), t is a t-norm, and f is a symmetrical function. The structuring element used in the dilatation δ_1 can simply be a unit ball, or a fuzzy set representing for instance the smallest sensitive unit in the image, along with the imprecision attached to it. In this case, it has to have a membership value equal to 1 at origin, in order to guarantee extensivity of dilations.

A distance density (Rosenfeld, 1985), i.e. a fuzzy number $D_N(\mu, \mu')(n)$ representing the degree to which the distance between μ and μ' is equal to n, can be obtained implicitly by:

$$\Delta_N(\mu, \mu')(n) = \int_0^n D_N(\mu, \mu')(n')dn'. \tag{14.59}$$

Clearly, this expression is not very tractable and does not lead to a simple explicit expression of $D_N(\mu, \mu')(n)$. Therefore, we suggest to use an explicit method, exploiting the other morphological expressions if nearest point distance (see Sec. 3.3). The translation of these equivalences provides, for $n > 0$, the following distance density:

(14.60)

$$D_N(\mu, \mu')(n) = t[\sup_{x \in \mathcal{S}} t[\mu'(x), \delta_n(\mu)(x)], c[\sup_{x \in \mathcal{S}} t[\mu'(x), \delta_{(n-1)}(\mu)(x)]]]$$

or a symmetrical expression derived from this one, and:

$$D_N(\mu, \mu')(0) = \sup_{x \in \mathcal{S}} t[\mu(x), \mu'(x)]. \tag{14.61}$$

Fuzzy Hausdorff distance. From Equation (14.37), a distance distribution can be defined, by introducing fuzzy dilation:

(14.62)
$$\Delta_H(\mu, \mu')(n) = t[\inf_{x \in \mathcal{S}} T[\delta_n(\mu)(x), c(\mu'(x))], \inf_{x \in \mathcal{S}} T[\delta_n(\mu')(x), c(\mu(x))]],$$

where c is a complementation, t a t-norm and T a t-conorm. A distance density can be derived implicitly from this distance distribution.

A direct definition of a distance density can be obtained from the expression of $H_d(X, Y) = n$ (see Sec. 3.3). Translating this expression leads to a definition of the Hausdorff distance between two fuzzy sets μ and μ' as a fuzzy number:

$$H_d(\mu, \mu')(0) = t[\inf_{x \in \mathcal{S}} T[\mu(x), c(\mu'(x))], \inf_{x \in \mathcal{S}} T[\mu'(x), c(\mu(x))]], \tag{14.63}$$

(14.64)
$$\begin{aligned} H_d(\mu, \mu')(n) = t[&\inf_{x \in \mathcal{S}} T[\delta_n(\mu)(x), c(\mu'(x))], \inf_{x \in \mathcal{S}} T[\delta_n(\mu')(x), c(\mu(x))], \\ &T(\sup_{x \in \mathcal{S}} t[\mu(x), c(\delta_{(n-1)}(\mu')(x))], \sup_{x \in \mathcal{S}} t[\mu'(x), c(\delta_{(n-1)}(\mu)(x))])]. \end{aligned}$$

Properties. These definitions of fuzzy nearest point and Hausdorff distances (defined as fuzzy numbers) between two fuzzy sets do not necessarily share the same properties as their crisp equivalent. All distances are positive, in the sense that the defined fuzzy numbers have always a support included in $\mathbb{R}^+$. By construction, all defined distances are symmetrical with respect to μ and μ'. The separability property (i.e. $d(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$) is not always satisfied. However, if μ is normalized (i.e. $\exists x, \mu(x) = 1$), we have for the nearest point distance $D_N(\mu, \mu)(0) = 1$ and $D_N(\mu, \mu)(n) = 0$ for $n > 1$. For the Hausdorff distance, $H_d(\mu, \mu')(0) = 1$ implies $\mu = \mu'$ for Lukasiewicz t-conorm, while it implies μ and μ' crisp and equal for $T = \max$. Also the triangular inequality is not satisfied in general.

Fuzzy adjacency from fuzzy dilation and set operations. Adjacency has a large interest in image processing, pattern recognition, spatial reasoning (Rosenfeld and Kak, 1976). A crisp definition of adjacency between crisp objects often leads to a low robustness, since the fact that two objects are adjacent or not may depend on one point only.

In order to account for possible errors or imprecisions, the framework of fuzzy sets is very useful. Two completely different ways for representing imprecision can be considered. In the first one, the satisfaction of the adjacency

property between two objects is considered to be a matter of degree; this can be more appropriate than a binary index (Rosenfeld, 1979; Rosenfeld, 1984). The second one consists in introducing imprecision in the objects themselves, and to deal with fuzzy objects. Then obviously adjacency is also a matter of degree. Only the second way is addressed here. More details can be found in Bloch et al., 1997.

Adjacency is defined using fuzzy dilation, by translating Equation (14.40) into fuzzy terms. The degree of adjacency between μ and ν involving fuzzy dilation is then:

$$\mu_{adj}(\mu,\nu) = t[\mu_{\neg int}(\mu,\nu), \mu_{int}[\delta_{B_c}(\mu),\nu], \mu_{int}[\delta_{B_c}(\nu),\mu]]. \tag{14.65}$$

This definition represents a conjunctive combination of a degree of non-intersection $\mu_{\neg int}$ between μ and ν and a degree of intersection μ_{int} between one fuzzy set and the dilation of the other. The degree of intersection can be defined using a supremum of a t-norm (as for fuzzy dilation):

$$\mu_{int}(\mu,\nu) = \sup_{x\in\mathcal{S}} t[\mu(x),\nu(x)], \tag{14.66}$$

or using the normalized fuzzy surface (or volume) of $t(\mu,\nu)$. The degree of non-intersection is simply defined by $\mu_{\neg int} = 1 - \mu_{int}$. B_c can be taken as the elementary structuring element related to the considered connectivity, or as a fuzzy structuring element, representing for instance spatial imprecision (i.e. the possibility distribution of the location of each point).

This degree of adjacency (with any structuring element) is symmetrical, consistent with the binary case and decreases when the distance between both fuzzy sets increases.

Fuzzy directional relative position from conditional fuzzy dilation. Relationships between objects can be partly described in terms of relative position, like "to the left of". Since such concepts are rather ambiguous, although human beings have an intuitive and common way of understanding and interpreting them, they may find a better modeling in the framework of fuzzy sets as fuzzy relationships. This framework makes it possible to propose flexible definitions which fit the intuition and may include subjective aspects, depending on the application and on the requirements of the user. Almost all existing methods for defining fuzzy relative directional spatial position (see Bloch and Ralescu, 2003 for a review) rely on angle measurements between points of the two objects of interest (Krishnapuram et al., 1993; Miyajima and Ralescu, 1994; Keller and Wang, 1995; Matsakis and Wendling, 1999), and concern 2D objects (sometimes with possible extension to 3D).

Another approach was proposed in Bloch, 1999a. It is based on fuzzy dilation and consists of two steps:

1 We first define a fuzzy "landscape" around the reference object R as a fuzzy set such that the membership value of each point corresponds to the degree of satisfaction of the spatial relation under examination. This fuzzy region is defined by a fuzzy dilation of the reference object by a fuzzy structuring element expressing the direction of interest $\vec{u}$ along with its imprecision. For instance, the structuring element ν can be defined as:

$$\forall P \in \mathcal{S},\ \nu(P) = f(\arccos \frac{\vec{OP} \cdot \vec{u}}{\|\vec{OP}\|}),\ \text{ and }\ \nu(O) = 1, \tag{14.67}$$

where O is the center of the structuring element and f is a decreasing function of $[0, \pi]$ into $[0, 1]$. An example is shown in Fig. 14.14. A fast algorithm for computing this fuzzy dilation is described in Bloch, 1999a.

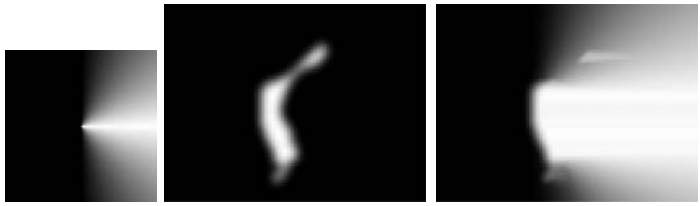


Figure 14.14. Fuzzy structuring element representing the relation "to the right of", a fuzzy reference object, and its dilation representing the region to the right of it (high grey values represent high membership values).

2 We then compare an object A to the fuzzy landscape attached to R, in order to evaluate how well this object matches with the areas having high membership values (i.e. areas that are in the desired direction). This is done using a fuzzy pattern matching approach (Dubois et al., 1988), which provides an evaluation as an interval instead of one number only. This makes a major difference with respect to all the previous approaches and, to our opinion, it provides a richer information about the considered relationship.

This definition is invariant with respect to translation, rotation and scaling, for 2D and 3D objects (crisp and fuzzy). When the distance between the objects increases, the objects are seen as points. The value of their relative position can be predicted only from the direction of interest and the direction in which one object goes far away from the reference object. Therefore the shape of the objects does no longer play any role in the assessment of their relative position. Finally, the behavior of the definition in cases where the reference object has strong concavities corresponds to what can be intuitively expected.

Reasoning on spatial relationships. Now, we address the second important issue in spatial reasoning. This includes fusion, since heterogeneous information has often to be combined in spatial reasoning, decision making and recognition (with a special focus on model-based recognition). Inference and logical reasoning are addressed in Sec. 4.

Fusion. Spatial reasoning aiming for instance at recognizing structures in an image has to deal with the combination of knowledge and information represented and modeled as described above. Usually, to achieve recognition, several spatial relationships to one or several spatial entities have to be combined, as well as information extracted from the image itself. For this combination step, the advantages of fuzzy sets lie in the variety of combination operators, which may deal with heterogeneous information expressed in a semi-quantitative framework (Dubois and Prade, 1985; Yager, 1991; Dubois et al., 1999). A classification of these operators with respect to their behavior (in terms of conjunctive, disjunctive, and compromise), the possible control of this behavior, their properties and their decisiveness proved to be useful for choosing an operator (Bloch, 1996).

Let us give a few examples. If we have different constraints about an object (for instance concerning the relations it should have with respect to another object) which have all to be satisfied, these constraints can be combined using a t-norm (a conjunction). This is typically the case when an object is described using relations to several objects or several relations of different types to the same object. If one object has to satisfy one relation or another one then a disjunction represented by a t-conorm has to be used. This occurs for instance when two symmetrical structures with respect to the reference object can be found. Mean operators can be used to combine several estimations and try to find a compromise between them.

Decision making and recognition. Let us now consider the introduction of fusion in model-based recognition procedures. We summarize here two distinct approaches. A first recognition approach, called global, uses the first type of question (1) raised at the beginning of Sec. 3.3 (define the degree to which a relation is satisfied between to given objects). The idea is to represent all available knowledge about the objects to be recognized. A typical example consists of graph-based representations. The model is then represented as a graph where nodes are objects and edges represent links between these objects. Both nodes and edges are attributed. Node attributes are characteristics of the objects, while edge attributes quantify spatial relationships between the objects. A data graph is then constructed from each image where the recognition has to be performed. Each region of the image (obtained after some processing) constitutes a node of this data graph, and edges represent spatial relationships

between regions, as for the model graph. The comparison between representations is performed through the computation of similarities between model graph attributes and data graph attributes. The fusion takes mainly place at this level, in order to combine the similarity values for different relationships. The fusion results constitute an objective function to be optimized by a matching procedure. This approach can benefit from the huge literature on fuzzy comparison tools (see e.g. Bouchon-Meunier et al., 1996) and from recent developments on fuzzy morphisms (Perchant and Bloch, 2002). It has been used in facial feature recognition based on a rough model of a face (Cesar et al., 2002) and brain structure recognition based on an anatomical atlas (Perchant et al., 1999; Bengoetxea et al., 2002).

A second type of approach relies on the second type of question (2) raised at the beginning of Sec. 3.3 (define the area of space in which a relation to a given reference object is satisfied), and is called here progressive. In such a progressive approach, objects are recognized sequentially and their recognition makes use of knowledge about their relations with respect to other objects. Relations with respect to previously obtained objects can be combined at two different levels of the procedure. First, fusion can occur in the spatial domain, using spatial fuzzy sets (Bloch et al., 2003). The result of this fusion allows to build a fuzzy region of interest in which the search of a new object will take place, in a process similar to focalization of attention. In a sequential procedure, the amount of available spatial relations increases with the number of processed objects. Therefore, the recognition of the most difficult structures, usually treated in the last steps, will be focused in a more restricted area. This approach has been used in medical imaging (Bloch et al., 2003; Colliot et al., 2004), as well as in mobile robotics to reason about the spatial position of the robot and the structure of its environment (Bloch and Saffiotti, 2002). Another fusion level occurs during the final decision step, i.e. segmentation and recognition of a structure. For this purpose, it was suggested in Colliot et al., 2004 to introduce relations in the evolution scheme of a deformable model, in which they are fused with other types of numerical information, usually edge and regularity constraints.

4. Logics

In this section, we explain how mathematical morphology relates to logics of space. As seen in Sec. 3.3, mathematical morphology can be considered as a spatial reasoning tool, with its two components: spatial knowledge representation and reasoning. Now we go one step further about the reasoning aspect, and we show how morphological operators can be applied on logical formulas (Sec. 4.1), and can be used to define a modal logic (Sec. 4.2). This leads to

qualitative representation of spatial relationships (Sec. 4.3), thus enhancing the power of logical reasoning with morphological aspects.

4.1 Morphology and propositional logics

We first consider the framework of propositional logics (note that this subsection is partly reproduced from Bloch and Lang, 2000).

In the knowledge representation community, propositional formulas are used to encode either pieces of knowledge (which may be generic—for instance, integrity constraints—or factual) or "preference items" (such as opinions, desires or goals), and are then used for complex reasoning or decision making tasks. These tasks often make use of operations on propositional formulas which are very similar to those considered in mathematical morphology. We give a (non-exhaustive) list of examples:

- *belief revision* (as shown by Katsuno and Mendelzon, 1991) consists of the following operation: let φ and ψ be two propositional formulas. The models of the revision $\varphi \circ \psi$ of φ by ψ are the models of ψ which are closest (with respect to a given distance) to a model of φ. Intuitively, using the language of mathematical morphology, it means that φ has to be dilated enough to intersect with some models of ψ. *Belief update* (Katsuno and Mendelzon, 1991) proceeds to the same kind of dilation but on each individual model of φ and then takes the union of all obtained sets of models.

- *belief merging* (Konieczny and Pino-Pérez, 1998) consists in finding the best compromise between a finite set of formulas $\varphi_1, \ldots, \varphi_n$, which amounts to selecting the models which minimize the aggregation (using some given operator) of the distances to each of the φ's. This amounts intuitively to dilate simultaneously all the φ's until they intersect. Similar operations are at work for the aggregation of preferences in group decision making as proposed in Lafage and Lang, 2000.

- one of the tasks involved in *similarity-based reasoning* (Esteva et al., 1997; Dubois et al., 1997) consists in determining if a formula φ approximatively entails a formula ψ by looking to what extent ψ has to be extended so as to contain all models of φ, which again corresponds to a dilation (and to directed Hausdorff distance).

- *reasoning with supermodels* (Ginsberg et al., 1998) uses models of a formula φ which are robust enough to resist some perturbations. In some cases, obtaining supermodels consists in eroding the formula so as to be far enough from the countermodels of φ. Again this corresponds to a classical operation of mathematical morphology (erosion). Another

close notion, evoked by Lafage and Lang, 2000, is the search for the most representative worlds of a formula.

- in *abductive reasoning* (Pino-Pérez and Uzcátegui, 1999), preferred explanations of a formula are defined based on a set of axioms, several of which being close to properties of morphological operators. Erosion appears as a useful tool in this context (Bloch et al., 2001; Bloch et al., 2004).

In this section, we investigate how and why mathematical morphology can be applied on logical formulas. First we note that the fact that a propositional formula can be equivalently defined by the set of its models enables us to apply easily all (set-theoretic) definitions of mathematical morphology to logical objects (worlds, formulas). This will lead us not only to rewriting well-known logical operations used for reasoning or decision making, but also to designing new kinds of logical objects or notions by transposing basic morphological operations to propositional logic. One may view morphological operators as transformations applied on formulas, leading to reasoning or decision making tools.

Basic logical concepts. Let PS be a finite set of propositional symbols. The set of formulas (generated by PS and the usual connectives) is denoted by Φ. Well-formed formulas will be denoted by Greek letters φ, ψ... Interpretations will be denoted by ω, ω' and the set of all interpretations for Φ by Ω. $Mod(\varphi) = \{\omega \in \Omega \mid \omega \models \varphi\}$ is the set of all models of φ (i.e. all interpretations for which φ is true).

The underlying idea for constructing morphological operations on logical formulas is to consider formulas and interpretations from a set theoretical perspective. Since Φ is isomorphic to 2^{Ω}, i.e., knowing a formula is equivalent to knowing the set of its models (and conversely, any set of models corresponds to a formula), we can identify φ with the set of its models $Mod(\varphi)$, and then apply set-theoretic morphological operations. We recall that $Mod(\varphi \vee \psi) = Mod(\varphi) \cup Mod(\psi)$, $Mod(\varphi \wedge \psi) = Mod(\varphi) \cap Mod(\psi)$, $Mod(\varphi) \subseteq Mod(\psi)$ iff $\varphi \models \psi$, and φ is consistent iff $Mod(\varphi) \neq \emptyset$.

Dilation and erosion of a formula. Using the previous equivalences, we propose to define dilation and erosion of a formula as follows:

$$Mod(\delta_B(\varphi)) = \{\omega \in \Omega \mid \check{B}_\omega \wedge \varphi \; consistent\}, \tag{14.68}$$

$$Mod(\varepsilon_B(\varphi)) = \{\omega \in \Omega \mid B_\omega \models \varphi\}. \tag{14.69}$$

In these equations, the structuring element B represents a relation between worlds, i.e. $\omega' \in B_\omega$ iff ω' satisfies some relationship with ω, and $\check{B}_\omega$ is defined

by $\omega' \in \check{B}_\omega \Leftrightarrow \omega \in B_{\omega'}$. The condition in Equation (14.68) expresses that the set of worlds in relation to ω should be consistent with φ. The condition in Equation (14.69) expresses that all worlds in relation to ω should be models of φ.

Structuring element. There are several possible ways to define structuring elements or more generally binary relations between worlds in a context of formulas. We suggest here a few ones. The relation between worlds defines a "neighbourhood" of worlds (equivalent to the neighbourhood function in Sec. 2.3). If it is symmetrical, it leads to symmetrical structuring elements. If it is reflexive, it leads to structuring elements such that $\omega \in B_\omega$, which leads to interesting properties, as will be seen later. For instance, this relationship can be an accessibility relation as in normal modal logics (Hughes and Cresswell, 1968). An interesting way to choose the relationship is to base it on distances between worlds. This allows to define sequences of increasing structuring elements defined as the balls of a distance. From any distance d between worlds, a distance from a world to a formula is derived as a distance from a point to a set: $d_N(\omega, \varphi) = \min_{\omega' \models \varphi} d(\omega, \omega')$.

The most commonly used distance between worlds in knowledge representation—especially in belief revision (Dalal, 1988), belief update (Katsuno and Mendelzon, 1991), merging (Konieczny and Pino-Pérez, 1998) or preference representation (Lafage and Lang, 2000)—is the Hamming distance d_H where $d_H(\omega, \omega')$ is the number of propositional symbols that are instantiated differently in both worlds. By default, we take d to be d_H.

Then dilation and erosion of size n are defined from Eqs. (14.68,14.69) by using the distance balls of radius n as structuring elements:

$$\begin{aligned} Mod(\delta_n(\varphi)) &= \{\omega \mid \exists \omega', \omega' \models \varphi \ and \ d_H(\omega, \omega') \leq n\} \\ &= \{\omega \mid d_N(\omega, \varphi) \leq n\}, \end{aligned} \tag{14.70}$$

$$\begin{aligned} Mod(\varepsilon_n(\varphi)) &= \{\omega \mid \forall \omega', d_H(\omega, \omega') \leq n \Rightarrow \omega' \models \varphi\} \\ &= \{\omega \mid d_N(\omega, \neg\varphi) > n\}. \end{aligned} \tag{14.71}$$

From operations with the unit ball we define the external (respectively internal) boundary of φ as $\delta_1(\varphi) \wedge \neg\varphi$ (respectively $\varphi \wedge \neg\varepsilon_1(\varphi)$), corresponding to the worlds that are exactly at distance 1 of φ (resp. of $\neg\varphi$).

Properties. The main properties of dilation and erosion, which are satisfied in mathematical morphology on sets, hold also in the logical setting proposed here, since the algebraic frameworks are the same up to an isomorphism.

Monotonicity: Both operators are increasing with respect to φ, i.e. if $\varphi \models \psi$, then $\delta_B(\varphi) \models \delta_B(\psi)$ and $\varepsilon_B(\varphi) \models \varepsilon_B(\psi)$, for any relation B. Dilation is increasing with respect to the relation, while erosion is decreasing, i.e. if $\forall \omega \in \Omega, B_\omega \subseteq B'_\omega$, then $\delta_B(\varphi) \models \delta_{B'}(\varphi)$ and $\varepsilon_{B'}(\varphi) \models \varepsilon_B(\varphi)$.

Extensivity: Dilation is extensive ($\varphi \models \delta_B(\varphi)$) if B is derived from a reflexive relation (as is the case for distance based dilation, since if $\omega \models \varphi$, then $d_N(\omega, \varphi) = 0$), and erosion is anti-extensive ($\varepsilon_B(\varphi) \models \varphi$) under the same conditions.

Iteration: Dilation and erosion satisfy an iteration property. For instance for distance based operations, we have:

$$\delta_{n+n'}(\varphi) = \delta_{n'}[\delta_n(\varphi)] = \delta_n[\delta_{n'}(\varphi)],$$

$$\varepsilon_{n+n'}(\varphi) = \varepsilon_{n'}[\varepsilon_n(\varphi)] = \varepsilon_n[\varepsilon_{n'}(\varphi)].$$

Commutativity with union or intersection: Dilation commutes with union or disjunction: for any family $\varphi_1, \ldots \varphi_m$ of formulas, we have: $\delta_B\big(\bigvee_{i=1}^m \varphi_i\big) = \bigvee_{i=1}^m \delta_B(\varphi_i)$. Erosion on the other hand commutes with intersection or conjunction.

In general dilation (resp. erosion) does not commute with intersection (resp. union), and only an inclusion relation holds: $\delta_B(\varphi \wedge \psi) \models \delta_B(\varphi) \wedge \delta_B(\psi)$.

Adjunction relation: $(\varepsilon_B, \delta_B)$ is an adjunction (moreover, if two operators form an adjunction, they are an erosion and a dilation respectively), i.e. $\delta_B(\psi) \models \varphi$ iff $\psi \models \varepsilon_B(\varphi)$.

Duality: Dilation and erosion (respectively opening and closing) are dual operators with respect to the negation: $\varepsilon_B(\varphi) = \neg\delta_{\check{B}}(\neg\varphi)$ which allows to deduce properties of an operator from those of its dual operator.

Relations to distances: Equation (14.70) is an example of how to derive a dilation from a distance. Conversely, we have: $d_N(\omega, \varphi) = \min\{n \in \mathbb{N} \mid \omega \models \delta_n(\varphi)\}$. Distances between formulas can also be derived from dilation, as minimum distance and Hausdorff distance. For instance the minimum distance (i.e. nearest world distance) is expressed as: $d_N(\varphi, \psi) = \min_{\omega \models \varphi, \omega' \models \psi} d_H(\omega, \omega') = \min\{n \in \mathbb{N} \mid \delta_n(\varphi) \wedge \psi \neq \emptyset \; and \; \delta_n(\psi) \wedge \varphi \neq \emptyset\}$. This means that the minimum distance is attained for the minimum size of dilation of each formula such that it becomes consistent with the other.

Opening and closing are defined classically by composition and have the same properties as the corresponding operators on sets. Filters, as described

in Sec. 2.4, can be applied on formulas as well, for instance for approximation and simplification purposes.

All these definitions and properties allow us to formalize problems of fusion, revision, abduction mentioned at the beginning of this subsection in morphological terms.

4.2 Morphological modal logic

When looking at the algebraic properties of mathematical morphology operators on the one hand, and of modal logic operators on the other hand, several similarities can be shown, and suggest that links between both theories are worth to be investigated. A pair of modal operators $(\Box, \Diamond)$ is defined in Bloch, 2002, as morphological erosion and dilation. This section summarizes this approach.

Until now mathematical morphology has been used mainly for quantitative and semi-quantitative (or semi-qualitative) representations of spatial relations. For qualitative spatial reasoning, several symbolic approaches have been developed, but mathematical morphology has not been widely used in this context. In Bloch, 2002, it was shown how modal operators based on morphological operators can be used for symbolic representations of spatial relations.

In a similar way as in Jeansoulin and Mathieu, 1994, the modal operators are used here for representing spatial relationships, and classical predicates represent the semantic part of the information. While inclusion and adjacency are considered in Jeansoulin and Mathieu, 1994, we consider here more spatial relationships, including metric ones, and model all of them using mathematical morphology.

Lattice structure. We use the same notations as in Sec. 4.1. We use standard Kripke's semantics and denote by $\mathcal{M}$ a model composed of a set of worlds Ω, a binary relation R between worlds and a truth valuation. Considering the inclusion relation on 2^Ω, $(2^\Omega, \subseteq)$ is a Boolean complete lattice. Similarly a lattice (which is isomorphic to 2^Ω) is defined on $\Phi_\equiv$, where $\Phi_\equiv$ denotes the quotient space of Φ by the equivalence relation between formulas (with the equivalence defined as $\varphi \equiv \psi$ iff $Mod(\varphi) = Mod(\psi)$). In the following, this is implicitly assumed, and we simply use the notation Φ. Any subset $\{\varphi_i\}$ of Φ has a supremum $\bigvee_i \varphi_i$, and an infimum $\bigwedge_i \varphi_i$ (corresponding respectively to union and intersection in 2^Ω). The greatest element is $\top$ and the smallest one is $\perp$ (corresponding respectively to 2^Ω and $\emptyset$). This lattice structure is important from the algebraic point of view of mathematical morphology. Indeed, it is the fundamental structure on which adjunctions and morphological operators can be defined.

A canonical formula φ_ω associated with a world ω is defined by:

$$Mod(\varphi_\omega) = \{\omega\}. \tag{14.72}$$

Let $\mathcal{C}$ be the subset of Φ containing all canonical formulas. The canonical formulas are sup-generating, i.e:

$$\forall \varphi \in \Phi,\ \exists \{\varphi_i\} \subseteq \mathcal{C},\ \varphi \equiv \bigvee_i \varphi_i. \tag{14.73}$$

The formulas φ_i are associated with the worlds ω_i which satisfy φ: for all ω_i such that $\omega_i \models \varphi$, $\varphi_i \equiv \varphi_{\omega_i}$. This decomposition is useful for some proofs.

Neighborhood function (or structuring element) as accessibility relation. The structuring element B representing a relationship between worlds defines a "neighbourhood" of worlds. This corresponds to the notion of neighbourhood function of Sec. 2.3. We propose to define this relationship as an accessibility relation as in normal modal logics (Hughes and Cresswell, 1968; Chellas, 1980).

An interesting way to choose the relationship is to base it on distances between worlds, as mentioned in Sec. 4.1. Another way to choose the relationship is to rely on an indistinguishability relation between worlds (Orlowska, 1993; Balbiani and Orlowska, 1999), for instance based on spatial attributes of spatial entities represented by these worlds. Interestingly enough, as shown in Orlowska, 1993, modal logics based on such relationships show some links with Pawlak's work on rough sets and rough logic (Pawlak, 1982; Pawlak, 1987), while rough sets can be constructed from morphological operators as shown in Bloch, 2000b. Also, the modal logic based on rough sets described e.g. in Yao and Lin, 1996, has links with the morphological modal logic described below. An accessibility relation can be defined from any neighbourhood function B as follows:

$$R(\omega, \omega') \text{ iff } \omega' \in B(\omega). \tag{14.74}$$

Conversely, a neighbourhood function can be defined from an accessibility relation using this equivalence. This is similar to the notions of Sec. 2.3.

The accessibility relation R is reflexive iff $\forall \omega \in \Omega,\ \omega \in B(\omega)$. It is symmetrical iff $\forall (\omega, \omega') \in \Omega^2,\ \omega \in B(\omega')$ iff $\omega' \in B(\omega)$. In general, accessibility relations derived from a neighbourhood function are not transitive. Indeed in general if $\omega' \in B(\omega)$ and $\omega'' \in B(\omega')$, we do not necessarily have $\omega'' \in B(\omega)$.

Modal logic from morphological dilations and erosions. Modal operators $\Box$ (necessity) and $\Diamond$ (possibility) are usually defined from an accessibility relation (Chellas, 1980) as:

$$\mathcal{M}, \omega \models \Box\varphi \text{ iff } \forall \omega' \in \Omega \text{ such that } R(\omega, \omega'), \mathcal{M}, \omega' \models \varphi, \tag{14.75}$$

$$\mathcal{M}, \omega \models \Diamond\varphi \text{ iff } \exists \omega' \in \Omega,\ R(\omega, \omega') \text{ and } \mathcal{M}, \omega' \models \varphi, \tag{14.76}$$

where $\mathcal{M}$ is a standard model related to R, that we will omit in the following in order to simplify notations (it will be always implicitly related to the considered accessibility relation).

Equation (14.75) can be rewritten as:

$$\begin{aligned} \omega \models \Box\varphi \quad &\Leftrightarrow \quad \{\omega' \in \Omega \mid R(\omega, \omega')\} \models \varphi \\ &\Leftrightarrow \quad \{\omega' \in \Omega \mid \omega' \in B(\omega)\} \models \varphi \\ &\Leftrightarrow \quad B(\omega) \models \varphi, \end{aligned}$$

which corresponds exactly to the definition of the erosion of a formula as defined in Equation (14.69).

Similarly, Equation (14.76) can be rewritten as:

$$\begin{aligned} \omega \models \Diamond\varphi \quad &\Leftrightarrow \quad \{\omega' \in \Omega \mid R(\omega, \omega')\} \cap Mod(\varphi) \neq \emptyset \\ &\Leftrightarrow \quad \{\omega' \in \Omega \mid \omega' \in B(\omega)\} \cap Mod(\varphi) \neq \emptyset \\ &\Leftrightarrow \quad B(\omega) \cap Mod(\varphi) \neq \emptyset, \end{aligned}$$

which exactly corresponds to a dilation according to Equation (14.68).

This shows that we can define modal operators based on an accessibility relation as erosion and dilation with a neighbourhood function:

$$\Box\varphi \equiv \varepsilon_B(\varphi), \tag{14.77}$$

$$\Diamond\varphi \equiv \delta_{\check{B}}(\varphi). \tag{14.78}$$

Let us now list the main properties of these operators. All results below can be found in Bloch, 2002, along with the corresponding proofs. Note that some results are direct consequences of the results of Sec. 2.3. For instance, **T** is deduced from Proposition 14.21, from which **5c** and **4c** are derived.

LEMMA 14.30 *The modal logic built from morphological erosions and dilations has the following theorems and rules of inference (we use similar notations as in Chellas, 1980):*

- **T**: $\Box\varphi \to \varphi$ *and* $\varphi \to \Diamond\varphi$ *iff* $\forall\omega \in \Omega, \omega \in B(\omega)$ *(reflexive accessibility relation).*
- **Df**: $\Diamond\varphi \leftrightarrow \neg\Box\neg\varphi$ *and* $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$.
- **D**: $\Box\varphi \to \Diamond\varphi$ *iff* R *is serial (or in other words,* $\forall\omega \in \Omega, B(\omega) \neq \emptyset$*).*
- **B**: $\Diamond\Box\varphi \to \varphi$ *and* $\varphi \to \Box\Diamond\varphi$ *for symmetrical* B.
- **5c**: $\Box\Diamond\varphi \to \Diamond\varphi$ *and* $\Box\varphi \to \Diamond\Box\varphi$ *iff* $\forall\omega \in \Omega, \omega \in B(\omega)$.
- **4c**: $\Box\Box\varphi \to \Box\varphi$ *and* $\Diamond\varphi \to \Diamond\Diamond\varphi$ *iff* $\forall\omega \in \Omega, \omega \in B(\omega)$.

- **N**: $\Box\top$ *and* $\neg\Diamond\bot$.
- **M**: $\Box(\varphi \wedge \psi) \rightarrow (\Box\varphi \wedge \Box\psi)$ *and* $(\Diamond\varphi \vee \Diamond\psi) \rightarrow \Diamond(\varphi \vee \psi)$.
- **M'**: $\Diamond(\varphi \wedge \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$ *and* $(\Box\varphi \vee \Box\psi) \rightarrow \Box(\varphi \vee \psi)$.
- **C**: $(\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \wedge \psi)$ *and* $\Diamond(\varphi \vee \psi) \rightarrow (\Diamond\varphi \vee \Diamond\psi)$.
- **R**: $(\Box\varphi \wedge \Box\psi) \leftrightarrow \Box(\varphi \wedge \psi)$ *and* $\Diamond(\varphi \vee \psi) \leftrightarrow (\Diamond\varphi \vee \Diamond\psi)$.
- **RN**:
$$\frac{\varphi}{\Box\varphi}.$$
- **RM**:
$$\frac{\varphi \rightarrow \psi}{\Box\varphi \rightarrow \Box\psi} \; and \; \frac{\varphi \rightarrow \psi}{\Diamond\varphi \rightarrow \Diamond\psi}.$$
- **RR**:
$$\frac{(\varphi \wedge \varphi') \rightarrow \psi}{(\Box\varphi \wedge \Box\varphi') \rightarrow \Box\psi} \; and \; \frac{(\varphi \vee \varphi') \rightarrow \psi}{(\Diamond\varphi \vee \Diamond\varphi') \rightarrow \Diamond\psi}.$$
- **RE**:
$$\frac{\varphi \leftrightarrow \psi}{\Box\varphi \leftrightarrow \Box\psi} \; and \; \frac{\varphi \leftrightarrow \psi}{\Diamond\varphi \leftrightarrow \Diamond\psi}.$$
- **K**: $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ *and by duality* $(\neg\Diamond\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\neg\varphi \wedge \psi)$.

Since the proposed system contains **Df**, **N**, **C** and is closed by **RM**, it is a normal modal logic (Chellas, 1980).

LEMMA 14.31 *On the contrary, the following expressions are not satisfied in general:*

- **5**: $\Diamond\varphi \rightarrow \Box\Diamond\varphi$ *(since for a symmetrical B the dilation followed by an erosion is a closing which does not necessarily contains the dilation).*
- **4**: $\Box\varphi \rightarrow \Box\Box\varphi$ *(since eroding a region twice produces a smaller region if $\omega \in B(\omega)$).*

Let us now denote by $\Box^n$ the iteration of n times $\Box$ (i.e. n erosions by the same structuring element). Since the succession of n erosions by a structuring element is equivalent to one erosion by a larger structuring element, of size n (iterativity property of erosion), $\Box^n$ is a new modal operator, constructed as in Equation (14.77). In a similar way, we denote by $\Diamond^n$ the iteration of n times $\Diamond$, which is again a new modal operator, due to iterativity property of dilation, constructed as in Equation (14.78) with a structuring element or neighbourhood of size n. We set $\Box^1 = \Box$ and $\Diamond^1 = \Diamond$.

We also have the following theorems for symmetrical B and $\omega \in B(\omega)$:

- $\Box^n \Box^{n'} \varphi \leftrightarrow \Box^{n+n'} \varphi$, and $\Diamond^n \Diamond^{n'} \varphi \leftrightarrow \Diamond^{n+n'} \varphi$ (iterativity properties of dilation and erosion). This property holds also in a more general case, without assumption on the symmetry of B.

- $\Diamond\Box\Diamond\Box\varphi \leftrightarrow \Diamond\Box\varphi$, and $\Box\Diamond\Box\Diamond\varphi \leftrightarrow \Box\Diamond\varphi$ (idempotence of opening and closing). This is actually a theorem from any KB logic: $\Diamond\Box\Diamond\Box\varphi \rightarrow \Diamond\Box\varphi$ is **B** applied to $\Diamond\Box\varphi$ and $\Diamond\Box\varphi \rightarrow \Diamond\Box\Diamond\Box\varphi$ comes from **B** applied to $\Box\varphi$ and from **RM**.

- $\Diamond\Box\Diamond\varphi \leftrightarrow \Diamond\varphi$, and $\Box\Diamond\Box\varphi \leftrightarrow \Box\varphi$.

- More generally, we derive from properties of opening and closing the following theorems:

$$\Diamond^n \Box^n \Diamond^{n'} \Box^{n'} \varphi \leftrightarrow \Diamond^{\max(n,n')} \Box^{\max(n,n')} \varphi,$$

  and

$$\Box^n \Diamond^n \Box^{n'} \Diamond^{n'} \varphi \leftrightarrow \Box^{\max(n,n')} \Diamond^{\max(n,n')} \varphi.$$

  See also the paragraph on granulometries in Sec. 2.4.

- For $n < n'$, the following expressions are theorems (if R is reflexive): $\Diamond^n \varphi \rightarrow \Diamond^{n'} \varphi$, $\Box^{n'} \varphi \rightarrow \Box^n \varphi$, $\Box^n \Diamond^n \varphi \rightarrow \Box^{n'} \Diamond^{n'} \varphi$, $\Diamond^{n'} \Box^{n'} \varphi \rightarrow \Diamond^n \Box^n \varphi$.

Modal operators from adjunction. Now, we consider the more general framework of algebraic erosions and dilations and the fundamental properties of adjunction (Heijmans and Ronse, 1990; Heijmans, 1994, Sec. 2.3).

Generalizing the definitions of Bloch and Lang, 2000, an algebraic dilation δ on Φ is defined as an operation which commutes with disjunction, and an algebraic erosion ε as an operation which commutes with conjunction, i.e. we have the two following expressions for any family $\{\varphi_i\}$:

$$\delta\Big(\bigvee_i \varphi_i\Big) \equiv \bigvee_i \delta(\varphi_i), \tag{14.79}$$

$$\varepsilon\Big(\bigwedge_i \varphi_i\Big) \equiv \bigwedge_i \varepsilon(\varphi_i). \tag{14.80}$$

One of the fundamental concept in the algebraic framework is the one of adjunction (see Sec. 2.3). Using similar concepts modal operators can be defined on Φ. A pair of modal operators $(\Box, \Diamond')$ is an adjunction on Φ iff:

$$\forall(\varphi, \psi) \in \Phi^2,\ \models (\Diamond'\varphi \rightarrow \psi \ \equiv\ \varphi \rightarrow \Box\psi), \tag{14.81}$$

or in other words:

$$\frac{\varphi \to \Box\psi}{\Diamond'\varphi \to \psi} \text{ and } \frac{\Diamond'\varphi \to \psi}{\varphi \to \Box\psi}.$$

In terms of worlds, this can also be expressed as:

(14.82) $\forall(\varphi,\psi) \in \Phi^2,\ Mod(\Diamond'\varphi) \subseteq Mod(\psi)$ iff $Mod(\varphi) \subseteq Mod(\Box\psi)$.

At this point, we use the notation $(\Box, \Diamond')$ instead of the classical notation $(\Box, \Diamond)$ because, as will be seen later, the two operators are not necessarily dual. In general they are two different modal operators.

LEMMA 14.32 *If $(\Box, \Diamond')$ is an adjunction on Φ, then $\Box$ is an algebraic erosion, and $\Diamond'$ is an algebraic dilation, i.e. for any family $\{\varphi_i\}$, we have:*

(14.83) $$\Box \bigwedge_i \varphi_i \equiv \bigwedge_i \Box\varphi_i,$$

(14.84) $$\Diamond' \bigvee_i \varphi_i \equiv \bigvee_i \Diamond'\varphi_i.$$

These equivalences are also true for empty families, since we have $\Box\top \equiv \top$ and $\Diamond'\bot \equiv \bot$.

LEMMA 14.33 *Let $(\Box, \Diamond')$ be an adjunction on the lattice of logical formulas. The modal logic based on these operators has the following theorems and rules of inference (we use similar notations as in Theorem 14.30 but $\Diamond$ has to be replaced by $\Diamond'$):* **B**, **N**, **M**, **M'**, **C**, **R**, **RN**, **RM**, **RR**, **RE**, **K**.

The proof is derived mainly from Theorem 14.32, from Eqs. (14.73, 14.81–14.84) and from the following result:

LEMMA 14.34 *We can write $\Box$ and $\Diamond'$ as:*

(14.85) $$\Box\varphi \equiv \bigvee\{\psi \in \Phi,\ \Diamond'\psi \to \varphi\},$$

(14.86) $$\Diamond'\varphi \equiv \bigwedge\{\psi \in \Phi,\ \varphi \to \Box\psi\}.$$

Again formulas are considered up to the equivalence relation, and therefore $\bigvee$ and $\bigwedge$ are taken over a finite family.

LEMMA 14.35 **T**, **5c** *and* **4c** *are not always satisfied, and we have the following results:*

- **T** *iff* $\forall\omega \in \Omega,\ \omega \models \Diamond'\varphi_\omega$,

- **5c** *iff* $\forall \omega \in \Omega,\ \omega \models \Diamond' \varphi_\omega$,
- **4c** *iff* $\forall \omega \in \Omega,\ \omega \models \Diamond' \varphi_\omega$.

Note that the condition on B for **T** in Theorem 14.30 corresponds to the one above, and we have $B(\omega) = Mod(\Diamond \varphi_\omega)$.

LEMMA 14.36 *We have the two following additional theorems:*

- $\Box\Diamond'\Box\varphi \leftrightarrow \Box\varphi$ *and* $\Diamond'\Box\Diamond'\varphi \leftrightarrow \Diamond'\varphi$.
- $\Diamond'\Box\Diamond'\Box\varphi \leftrightarrow \Diamond'\Box\varphi$ *and* $\Box\Diamond'\Box\Diamond'\varphi \leftrightarrow \Box\Diamond'\varphi$.

LEMMA 14.37 *Let* $(\Box, \Diamond')$ *be an adjunction on* Φ. *Let* $\Box_*\varphi \equiv \neg\Box\neg\varphi$ *and* $\Diamond'_*\varphi \equiv \neg\Diamond'\neg\varphi$. *Then* $(\Diamond'_*, \Box_*)$ *is an adjunction.*

This property expresses a kind of duality between both operators.

Note that we do not always have: **Df**: $\Diamond'\varphi \leftrightarrow \neg\Box\neg\varphi$ and $\Box\varphi \leftrightarrow \neg\Diamond'\neg\varphi$, nor **D**: $\Box\varphi \rightarrow \Diamond'\varphi$.

LEMMA 14.38 **Df** *is satisfied by an adjunction* $(\Box, \Diamond')$ *if and only if* $\Diamond'$ *satisfies the following property:*

$$\forall(\omega, \omega') \in \Omega^2,\ \omega \models \Diamond'\varphi_{\omega'} \text{ iff } \omega' \models \Diamond'\varphi_\omega. \tag{14.87}$$

D *is satisfied by an adjunction* $(\Box, \Diamond')$ *if* $\Diamond'$ *satisfies one of the two following properties:*

$$\forall \omega \in \Omega,\ \omega \models \Diamond'\varphi_\omega \tag{14.88}$$

or

(14.89)

$$\forall(\omega, \omega') \in \Omega^2,\ \omega \models \Diamond'\varphi_{\omega'} \text{ iff } \omega' \models \Diamond'\varphi_\omega \text{ and } \{\omega', \omega' \models \Diamond'\varphi_\omega\} \neq \emptyset.$$

The last result (see Proposition 14.20) means in particular that we can have **D** without having **T**.

In cases where **Df** is satisfied, then we note simply $\Diamond$ instead of $\Diamond'$.

LEMMA 14.39 *The operators* $(\Box, \Diamond)$ *defined by Eqs. (14.77,14.78) build an adjunction in the case* B *is symmetrical.*

This shows that modal operators derived from morphological erosions and dilations are particular cases of modal operators derived from algebraic erosions and dilations.

All these results show that the use of general algebraic dilations and erosions defined from the adjunction property lead to the properties of normal modal logics. This justifies the use of Kripke's semantics. This also guarantees a completeness result.

Characterizing modal logics in terms of morphological operators. Conversely, the following result shows that modal operators satisfying some axioms can be expressed in morphological terms.

LEMMA 14.40 *If two modal operators $\Box$ and $\Diamond$ satisfy* **B** *and* **RM**, *then $(\Box, \Diamond)$ is an adjunction on Φ, $\Box$ is an algebraic erosion and $\Diamond$ is an algebraic dilation.*

Moreover, if we define a relation R between worlds by $R(\omega, \omega')$ iff $\omega \models \Diamond\varphi_{\omega'}$, where φ_ω is a canonical formula associated with ω ($Mod(\varphi_\omega) = \{\omega\}$), then $\Box$ and $\Diamond$ are exactly given by:

$$Mod(\Box\varphi) = \{\omega \in \Omega \mid \forall\omega', R(\omega', \omega) \Rightarrow \omega' \models \varphi\}, \tag{14.90}$$

$$Mod(\Diamond\varphi) = \{\omega \in \Omega \mid \exists\omega', R(\omega, \omega'), \omega' \models \varphi\}. \tag{14.91}$$

These equations are similar to the ones used for defining modal operators from an accessibility relation and a structuring element, except that here we consider $R(\omega, \omega')$ for one operator, and $R(\omega', \omega)$ for the other. If R is symmetrical, both are equivalent. In cases where the structuring element (and the accessibility relation) is not symmetrical, we consider its symmetrical in one of the operations.

Modal operators from morphological opening and closing. We can define modal operators from opening and closing on formulas as:

$$\Box\varphi \equiv O(\varphi), \tag{14.92}$$

$$\Diamond\varphi \equiv C(\varphi). \tag{14.93}$$

Unfortunately, this leads to weaker properties than operators derived from erosion and dilation. This comes partly from the fact that no accessibility relation can be derived from opening and closing as easily as from erosion and dilation.

However, it would be interesting to link this approach with the topological interpretation of modal logic as proposed in Aiello and van Benthem, 1999, since opening and closing are related to the notions of topological interior and closure. Note that considering erosion and dilation only leads to a pre-topology (where closure is not idempotent).

Another interesting direction could be to consider the neighbourhood semantics (Aiello and van Benthem, 1999), where here the neighbourhoods of ω would be all elements of the set $N(\omega) = \{B(\omega') \mid \omega' \in \Omega \text{ and } \omega \in B(\omega')\}$. With this semantics, we can prove:

$$\omega \models \Box\varphi \Leftrightarrow \exists\omega' \in \Omega \mid B(\omega') \in N(\omega) \text{ and } B(\omega') \models \varphi. \tag{14.94}$$

The proof of this expression comes from the following rewriting of opening:

$$Mod(\Box\varphi) = \{\omega \in \Omega \mid \exists\omega' \in \Omega,\ \omega \in B(\omega') \text{ and } B(\omega') \models \varphi\}. \tag{14.95}$$

Kripke's semantics can be seen as a particular case, where the neighbourhood of ω is reduced to the singleton $\{B(\omega)\}$.

LEMMA 14.41 *The modal logic constructed from opening and closing satisfies* **T**, **Df**, **D**, **4**, **4c**, **5c**, **N**, **M**, **M'**, **RM**, **RE**, *but not* **5**, **B**, **K**, **C**, **R**, **RR**.

The fact that **K** is not satisfied goes with the interpretation in terms of neighbourhood semantics, which leads to a weaker logic, where **RM** (monotonicity) is satisfied, but not **K** in general (Aiello and van Benthem, 1999).

Extension to the fuzzy case. We now consider fuzzy formulas, i.e. formulas φ for which $Mod(\varphi)$ is a fuzzy subset of Ω and use the fuzzy morphological operators of Bloch and Maître, 1995, (see Sec. 3.4). However, what follows applies as well if other definitions are used.

Modal operators in the fuzzy case can then be constructed from fuzzy erosion and dilation in a similar way as in the crisp case using Eqs. (14.77, 14.78). The fuzzy structuring element can be interpreted as a fuzzy relation between worlds. The properties of this fuzzy modal logic are the same as in the crisp case, since fuzzy dilations and erosions have the same properties as the binary ones.

This extension can also be considered from the algebraic point of view of adjunction, based on the results of Deng and Heijmans, 2002, and on a definition of fuzzy erosion in terms of residual implication.

The use of fuzzy structuring elements will appear as particularly useful for expressing intrinsically vague spatial relationships such as directional relative position.

It is also interesting to relate this approach to the possibilistic logic proposed for belief fusion in Boldrin and Saffiotti, 1995, and to similarity-based reasoning (Esteva et al., 1997; Dubois et al., 1997).

4.3 Qualitative representation of spatial relationships and reasoning

For qualitative spatial reasoning, worlds (or interpretations) can represent spatial entities, like regions of the space. Formulas then represent combinations of such entities, and define regions, objects, etc., which may be not connected. For instance, if a formula φ is a symbolic representation of a region X of the space, it can be interpreted for instance as "the object we are looking at is in X". In an epistemic interpretation, it could represent the belief of an agent that the object is in X. The interest of such representations could be also to deal in a qualitative way with any kind of spatial entities, without referring to points.

Using these interpretations, if φ represents some knowledge or belief about a region X of the space, then $\Box\varphi$ represents a restriction of X. If we are looking at an object in X, then $\Box\varphi$ is a necessary region for this object. Similarly, $\Diamond\varphi$ represents an extension of X, and a possible region for the object. In an epistemic interpretation, $\Box\varphi$ can represent the belief of an agent that the object is necessarily in the erosion of X while $\Diamond\varphi$ is the belief that it is possibly in the dilation of X. Interpretations in terms of rough regions are also possible.

In this subsection, we address the problem of qualitative representation of spatial relationships between regions or objects represented by logical formulas.

Topological relationships. Let us first consider topological relationships. Let φ and ψ be two formulas representing two regions X and Y of the space. Note that all what follows holds in both crisp and fuzzy cases. Simple topological relations such as inclusion, exclusion, intersection do not call for more operators than the standard ones of propositional logic (e.g. Bennett, 1995). But other relations such that X *is a tangential part of* Y can benefit from the morphological modal operators. Such a relationship can be expressed as:

(14.96) $$\varphi \to \psi \text{ and } \Diamond\varphi \wedge \neg\psi \text{ consistent,}$$

or, equivalently,

(14.97) $$\varphi \to \psi \text{ and } \varphi \wedge \neg\Box\psi \text{ consistent.}$$

Indeed, if X is a tangential part of Y, it is included in Y but its dilation is not, and equivalently it is not included in the erosion of Y, as illustrated in Fig. 14.13.

In a similar way, a relation such that X is a non tangential part of Y is expressed, for a reflexive accessibility relation, as:

(14.98) $$\Diamond\varphi \to \psi,$$

or, equivalently,

(14.99) $$\varphi \to \Box\psi,$$

(i.e. in order to verify that X is a non tangential part of Y, we have to prove these relations).

If we also want X to be a proper part, we have to add the following condition:

(14.100) $$\neg\varphi \wedge \psi \text{ consistent.}$$

Let us now consider adjacency (or external connection). Saying that X is adjacent to Y means that they do not intersect and as soon as one region is

dilated, it has a non empty intersection with the other one. In symbolic terms, this relation can be expressed as:

(14.101) $\varphi \wedge \psi$ inconsistent and $\Diamond\varphi \wedge \psi$ consistent (or $\varphi \wedge \Diamond\psi$ consistent).

Actually, this expression holds in a discrete domain. If φ and ψ represent spatial entities in a continuous spatial domain, some problems may occur if these entities are closed sets and have parts of local dimension less than the dimension of the space (see Bloch et al., 1997 for a complete discussion). Such problems can be avoided if the entities are reduced to regular ones, i.e. that are equal to the closure of their interior (and by considering an asymptotic definition of adjacency). Using the topological interpretation of modal logic, this amounts to deal with formulas for which we can prove $\varphi \leftrightarrow \Diamond\Box\varphi$.

It is interesting to link these types of representations with the ones developed in the community of mereology and mereotopology, where such relations are defined respectively from parthood and connection predicates (Asher and Vieu, 1995; Randell et al., 1992; Cohn et al., 1997; Varzi, 1996; Renz and Nebel, 2001). Interestingly enough, erosion is defined from inclusion (i.e. a parthood relationship) and dilation from intersection (i.e. a connection relationship). Some axioms of these domains could be expressed in terms of dilation. For instance from a parthood postulate $P(X, Y)$ between two spatial entities X and Y and from dilation δ, tangential proper part could be defined as $TPP(X, Y) = P(X, Y) \wedge \neg P(Y, X) \wedge \neg P(\delta(X), Y)$. Further links certainly deserve to be investigated, in particular with the work presented e.g. in Cohn et al., 1997, Cristani et al., 2000 and Galton, 2000.

Distances. Distances between objects X and Y can be expressed in different forms, as *the distance between X and Y is equal to n, the distance between X and Y is less (respectively greater) than n, the distance between X and Y is between n_1 and n_2*. Several distances can be related to morphological dilation, as minimum distance and Hausdorff distance, as explained in Sec. 3.3.

Based on algebraic expressions of distances using dilation, the translation into a logical formalism is quite straightforward. Expressing that $d_N(X, Y) = n$ leads to:

(14.102) $$\begin{cases} \forall m < n, \Diamond^m\varphi \wedge \psi \text{ inconsistent, and } \Diamond^m\psi \wedge \varphi \text{ inconsistent} \\ \text{and } \Diamond^n\varphi \wedge \psi \text{ consistent (or } \Diamond^n\psi \wedge \varphi \text{ consistent).} \end{cases}$$

Expressions like $d_N(X, Y) \leq n$ translate into:

(14.103) $$\Diamond^n\varphi \wedge \psi \text{ consistent (or } \Diamond^n\psi \wedge \varphi \text{ consistent).}$$

Expressions like $d_N(X, Y) \geq n$ translate into:

(14.104) $$\forall m < n, \Diamond^m\varphi \wedge \psi \text{ inconsistent (or } \Diamond^m\psi \wedge \varphi \text{ inconsistent).}$$

Expressions like $n_1 \leq d_N(X,Y) \leq n_2$ translate into:

$$\text{(14.105)} \quad \begin{cases} \forall m < n_1, \Diamond^m \varphi \wedge \psi \text{ inconsistent (or } \Diamond^m \psi \wedge \varphi \text{ inconsistent)} \\ \text{and } \Diamond^{n_2} \varphi \wedge \psi \text{ consistent (or } \Diamond^{n_2} \psi \wedge \varphi \text{ consistent).} \end{cases}$$

Similarly for Hausdorff distance, we translate $H_d(X,Y) = n$ by:

$$\text{(14.106)} \quad \begin{cases} \forall m < n, \psi \wedge \neg\Diamond^m \varphi \text{ consistent or } \varphi \wedge \neg\Diamond^m \psi \text{ consistent} \\ \text{and } \psi \rightarrow \Diamond^n \varphi \text{ and } \varphi \rightarrow \Diamond^n \psi. \end{cases}$$

The first condition corresponds to $H_d(X,Y) \geq n$ and the second one to $H_d(X,Y) \leq n$.

Let us consider an example of possible use of these representations for spatial reasoning. If we are looking at an object represented by ψ in an area which is at a distance in an interval $[n_1, n_2]$ of a region represented by φ, this corresponds to a minimum distance greater than n_1 and to a Hausdorff distance less than n_2. This is illustrated in Fig. 14.12.

Then we have to check the following relation:

$$\text{(14.107)} \qquad \psi \rightarrow \neg\Diamond^{n_1} \varphi \wedge \Diamond^{n_2} \varphi,$$

or equivalently:

$$\text{(14.108)} \qquad \psi \rightarrow \Box^{n_1} \neg\varphi \wedge \Diamond^{n_2} \varphi.$$

This expresses in a symbolic way an imprecise knowledge about distances represented as an interval. If we consider a fuzzy interval, this extends directly by means of fuzzy dilation (see Bloch, 2000c, for detailed expressions of these dilations).

These expressions show how we can convert distance information, which is usually defined in an analytical way, into algebraic expressions through mathematical morphology, and then into logical expressions through morphological expressions of modal operators.

Directional relative position. We use for this relation the same approach as in Sec. 3.4, based on dilation. Let us denote by D^d the dilation corresponding to a directional information in the direction d, and by $\Diamond^d$ the associated modal operator (this assumes that directions can be defined over the set of spatial entities represented as logical formulas). Expressing that an object represented by ψ has to be in direction d with respect to a region represented by φ amounts to check the following relation:

$$\text{(14.109)} \qquad \psi \rightarrow \Diamond^d \varphi.$$

In the fuzzy case, this relation can hold to some degree.

Usually for spatial reasoning several relationships have to be used together. This aspect can benefit from the developments in information fusion, both in a numerical and in a logical setting.

Logical reasoning and inference. One of the advantages of logical representations is their inference and reasoning power. Rule-based systems can make use of the proposed representations in a quite straightforward way. But it is also interesting to note that several spatial logics contain ingredients that can be expressed equivalently in morphological terms. We show here some of these links but do not pretend to be exhaustive.

Some links with mereotopology and region connection calculus (RCC) have already been mentioned above. They allow us to combine the expressiveness power of mathematical morphology and the reasoning power of RCC and mereotopology.

The "egg-yolk" structures, as developed e.g. in Cristani et al., 2000 can also lead to interpretations in terms of mathematical morphology. For instance in this model, establishing if a yolk can be a mobile part (in translation) of its egg is based on the notion of congruence. This characterization can be expressed in a very simple way using morphological opening (erosion followed by a dilation): the opening of the egg by the yolk considered as the structuring element should be connected.

Let us now consider two examples of logics of distances. The first one defines a modality $A^{\leq a}$ by Kutz et al., 2002:

$$\omega \models A^{\leq a}\varphi \text{ iff } \forall u, d(\omega, u) \leq a \Rightarrow u \models \varphi, \tag{14.110}$$

where d is a distance between worlds. It is straightforward to show that $A^{\leq a}\varphi$ is equivalent to the erosion of φ by a ball of the distance d of radius a. The dual of $A^{\leq a}$ is equivalent to a dilation. Then we have direct correspondences between the axioms of this distance logics and the axioms of the modal morpho-logics as presented in Bloch, 2002. Some theorems can be also directly deduced from properties of dilation or erosion. For instance, the following is proved to be a theorem:

$$A^{\leq b}\varphi \rightarrow A^{\leq a}\varphi \text{ for } a \leq b. \tag{14.111}$$

Using the morphological equivalence, this theorem is directly deduced from the decreasingness of erosion with respect to the size of the structuring element.

The second example concerns nearness logics (Aiello and van Benthem, 2002), where the notion of "nearest than" is modeled as:

$$x \models < N > \varphi, \psi \text{ iff } \exists y, z, (y \models \varphi \wedge z \models \psi) \wedge N(x, y, z) \tag{14.112}$$

where $N(x, y, z)$ means that y is closer to x than z is. The meaning of this expression is that the nearest point distance of x to φ is less than the nearest point distance of x to ψ. An equivalent expression is therefore:

$$x \models \delta_n(\psi) \rightarrow x \models \delta_n(\varphi) \tag{14.113}$$

which expresses that x is reached faster from φ than from ψ by dilations of these formulas.

Other links between linear logics or arrow logics and mathematical morphology exist, as already established in Aiello and van Benthem, 2002.

Finally, let us consider logics of convexity (Aiello and van Benthem, 2002):

$$x \models C\varphi \text{ iff } \exists y, z, (y \models \varphi \wedge z \models \varphi) \wedge (x \in y - z) \tag{14.114}$$

which expresses a linear closure, the iteration of which provides convexity. This iterative closure is clearly equivalent to morphologic closing, where structuring elements are segments in all directions of infinite length (in practice, larger than the largest diameter of the considered spatial entities).

All these examples show interesting links between different spatial logics which have not been exhibited before for most of them. They can be exploited in two ways: the properties of morphological operators can provide additional theorems to these logics; conversely spatial logics endow mathematical morphology with powerful inference and reasoning tools.

Other links between mathematical morphology and non-classical logics are explored in Fujio and Bloch, 2004.

5. Conclusion

This chapter provides an overview of the algebraic basis of mathematical morphology. It shows how this lattice-theoretical formalism can be applied in different frameworks for spatial reasoning, thanks to the representation of shapes and of spatial relations it provides. In particular, it is highly adapted to the modelling of logical relations.

Further links with other chapters in this book are worth to be mentioned, such as mereotopology, modal logics of space and topology.

It should be stressed that we have not given here a complete overview of the methods, techniques and tools of mathematical morphology. Nor have we given any idea of the way to use them (alone or in conjunction with other approaches) in practical image processing problems. This by far exceeds the scope of this chapter. For this, we advise reading classical books such as Heijmans, 1994, Serra, 1982, Serra, 1988 and Soille, 2003, as well as image processing journals and conference proceedings.

Complexity issues have not been addressed in this chapter, except a few words at the end of Sec. 1. The interested reader should consult standard books on morphological image processing (for example, Soille, 2003) for more details on algorithms and data structures for morphology.

References

Aiello, M. (2002). *Spatial Reasoning, Theory and Practice*. PhD thesis, University of Amsterdam.

Aiello, M. and van Benthem, J. (1999). Logical Patterns in Space. Technical Report UVA PP-99-24, University of Amsterdam.

Aiello, M. and van Benthem, J. (2002). A Modal Walk Through Space. *Journal of Applied Non Classical Logics*, 12(3-4):319–364.

Alexandroff, P. (1937). Diskrete Raüme. *Mat. Sb.*, 2:501–518.

Alexandroff, P. and Hopf, H. (1935). *Topologie, Erster Band*. Springer-Verlag, Berlin.

Alexandroff, P. S. (1956). *Combinatorial Topology, Vol. 1*. Graylock Press, Rochester, NY.

Asher, N. and Vieu, L. (1995). Toward a Geometry of Common Sense: A Semantics and a Complete Axiomatization of Mereotopology. In *IJCAI'95*, pages 846–852, San Mateo, CA.

Balbiani, P. and Orlowska, E. (1999). A Hierarchy of Modal Logics with Relative Accessibility Relations. *Journal of Applied Non-Classical Logics*, 9: 303–328.

Bandemer, H. and Näther, W. (1992). *Fuzzy Data Analysis*. Theory and Decision Library, Serie B: Mathematical and Statistical Methods. Kluwer Academic Publisher, Dordrecht.

Bengoetxea, E., Larranaga, P., Bloch, I., Perchant, A., and Boeres, C. (2002). Inexact Graph Matching by Means of Estimation of Distribution Algorithms. *Pattern Recognition*, 35:2867–2880.

Bennett, B. (1995). Modal Logics for Qualitative Spatial Reasoning. *Bulletin of the IGPL*, 4(1):23–45.

Birkhoff, G. (1995). *Lattice Theory*. Vol. 25 of American Mathematical Society Colloquium Publications. American Mathematical Society, 3rd edition, 8th printing.

Bloch, I. (1993). Triangular Norms as a Tool for Constructing Fuzzy Mathematical Morphologies. In *International Workshop on "Mathematical Morphology and its Applications to Signal Processing"*, pages 157–161, Barcelona, Spain.

Bloch, I. (1996). Information Combination Operators for Data Fusion: A Comparative Review with Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(1):52–67.

Bloch, I. (1999a). Fuzzy Relative Position between Objects in Image Processing: a Morphological Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):657–664.

Bloch, I. (1999b). On Fuzzy Distances and their Use in Image Processing under Imprecision. *Pattern Recognition*, 32(11):1873–1895.

Bloch, I. (2000a). Geodesic Balls in a Fuzzy Set and Fuzzy Geodesic Mathematical Morphology. *Pattern Recognition*, 33(6):897–905.

Bloch, I. (2000b). On Links between Mathematical Morphology and Rough Sets. *Pattern Recognition*, 33(9):1487–1496.

Bloch, I. (2000c). Spatial Representation of Spatial Relationships Knowledge. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *7th International Conference on Principles of Knowledge Representation and Reasoning KR 2000*, pages 247–258, Breckenridge, CO. Morgan Kaufmann, San Francisco, CA.

Bloch, I. (2002). Modal Logics based on Mathematical Morphology for Spatial Reasoning. *Journal of Applied Non Classical Logics*, 12(3-4):399–424.

Bloch, I. (2003a). On Fuzzy Spatial Distances. In Hawkes, P., editor, *Advances in Imaging and Electron Physics, Volume 128*, pages 51–122. Elsevier, Amsterdam.

Bloch, I. (2003b). Unifying Quantitative, Semi-Quantitative and Qualitative Spatial Relation Knowledge Representations using Mathematical Morphology. In Asano, T., Klette, R., and Ronse, C., editors, *LNCS 2616 Geometry, Morphology, and Computational Imaging, 11th International Workshop on Theoretical Foundations of Computer Vision, Dagstuhl Castle, Germany, April 7–12, 2002, Revised Papers*, pages 153–164. Springer.

Bloch, I., Géraud, T., and Maître, H. (2003). Representation and Fusion of Heterogeneous Fuzzy Information in the 3D Space for Model-Based Structural Recognition - Application to 3D Brain Imaging. *Artificial Intelligence*, 148:141–175.

Bloch, I. and Lang, J. (2000). Towards Mathematical Morpho-Logics. In *8th International Conference on Information Processing and Management of Uncertainty in Knowledge based Systems IPMU 2000*, volume III, pages 1405–1412, Madrid, Spain.

Bloch, I. and Maître, H. (1995). Fuzzy Mathematical Morphologies: A Comparative Study. *Pattern Recognition*, 28(9):1341–1387.

Bloch, I., Maître, H., and Anvari, M. (1997). Fuzzy Adjacency between Image Objects. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5(6):615–653.

Bloch, I., Pino-Pérez, R., and Uzcátegui, C. (2001). Explanatory Relations based on Mathematical Morphology. In *ECSQARU 2001*, pages 736–747, Toulouse, France.

Bloch, I., Pino-Pérez, R., and Uzcategui, C. (2004). A Unified Treatment of Knowledge Dynamics. In *International Conference on the Principles of Knowledge Representation and Reasoning, KR2004*, pages 329–337, Canada.

Bloch, I. and Ralescu, A. (2003). Directional Relative Position between Objects in Image Processing: A Comparison between Fuzzy Approaches. *Pattern Recognition*, 36:1563–1582.

Bloch, I. and Saffiotti, A. (2002). On the Representation of Fuzzy Spatial Relations in Robot Maps. In *IPMU 2002*, volume III, pages 1587–1594, Annecy, France.

Boldrin, L. and Saffiotti, A. (1995). A Modal logic for merging Partial Belief of Multiple Reasoners. Technical Report TR/IRIDIA/95-19, IRIDIA, Université Libre de Bruxelles.

Bouchon-Meunier, B., Rifqi, M., and Bothorel, S. (1996). Towards General Measures of Comparison of Objects. *Fuzzy Sets and Systems*, 84(2): 143–153.

Cesar, R., Bengoetxea, E., and Bloch, I. (2002). Inexact Graph Matching using Stochastic Optimization Techniques for Facial Feature Recognition. In *International Conference on Pattern Recognition ICPR 2002*, volume 2, pages 465–468, Québec.

Chellas, B. (1980). *Modal Logic, an Introduction*. Cambridge University Press, Cambridge.

Cohn, A., Bennett, B., Gooday, J., and Gotts, N. M. (1997). Representing and Reasoning with Qualitative Spatial Relations about Regions. In Stock, O., editor, *Spatial and Temporal Reasoning*, pages 97–134. Kluwer.

Colliot, O., Camara, O., Dewynter, R., and Bloch, I. (2004). Description of Brain Internal Structures by Means of Spatial Relations for MR Image Segmentation. In *SPIE Medical Imaging*, volume 5370, pages 444–455, San Diego, CA, USA.

Cristani, M., Cohn, A. G., and Bennett, B. (2000). Spatial Locations via Morpho-Mereology. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *7th International Conference on Principles of Knowledge Representation and Reasoning KR 2000*, pages 15–25, Breckenridge, CO. Morgan Kaufmann, San Francisco, CA.

Dalal, M. (1988). Investigations into a Theory of Knowledge Base Revision: Preliminary Report. In *AAAI'88*, pages 475–479.

de Baets, B. (1995). Idempotent Closing and Opening Operations in Fuzzy Mathematical Morphology. In *ISUMA-NAFIPS'95*, pages 228–233, College Park, MD.

de Baets, B. (1997). Fuzzy Morphology: a Logical Approach. In Ayyub, B. and Gupta, M., editors, *Uncertainty in Engineering and Sciences: Fuzzy Logic, Statistics and Neural Network Approach*, pages 53–67. Kluwer Academic.

de Baets, B. and Kerre, E. (1993). An Introduction to Fuzzy Mathematical Morphology. In *NAFIPS'93*, pages 129–133, Allentown, Pennsylvania.

Deng, T.-Q. and Heijmans, H. (2002). Grey-Scale Morphology Based on Fuzzy Logic. *Journal of Mathematical Imaging and Vision*, 16:155–171.

di Gesu, V. (1988). Mathematical Morphology and Image Analysis: A Fuzzy Approach. In *Workshop on Knowledge-Based Systems and Models of Logical Reasoning*, Reasoning.

di Gesu, V., Maccarone, M. C., and Tripiciano, M. (1993). Mathematical Morphology based on Fuzzy Operators. In Lowen, R. and Roubens, M., editors, *Fuzzy Logic*, pages 477–486. Kluwer Academic.

Dubois, D., Esteva, F., Garcia, P., Godo, L., and Prade, H. (1997). A Logical Approach to Interpolation based on Similarity Relations. *International Journal of Approximate Reasoning*, 17(1):1–36.

Dubois, D. and Jaulent, M.-C. (1987). A General Approach to Parameter Evaluation in Fuzzy Digital Pictures. *Pattern Recognition Letters*, 6:251–259.

Dubois, D. and Prade, H. (1980). *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, New-York.

Dubois, D. and Prade, H. (1983). Inverse Operations for Fuzzy Numbers. In Sanchez, E. and Gupta, M., editors, *Fuzzy Information, Knowledge Representation and Decision Analysis, IFAC Symposium*, pages 391–396, Marseille, France.

Dubois, D. and Prade, H. (1985). A Review of Fuzzy Set Aggregation Connectives. *Information Sciences*, 36:85–121.

Dubois, D. and Prade, H. (1990). Rough Fuzzy Sets and Fuzzy Rough Sets. *International Journal of General Systems*, 17:191–209.

Dubois, D. and Prade, H. (1991). Fuzzy Sets in Approximate Reasoning, Part I: Inference with Possibility Distributions. *Fuzzy Sets and Systems*, 40: 143–202.

Dubois, D., Prade, H., and Testemale, C. (1988). Weighted Fuzzy Pattern Matching. *Fuzzy Sets and Systems*, 28:313–331.

Dubois, D., Prade, H., and Yager, R. (1999). Merging Fuzzy Information. In Bezdek, J.C., Dubois, D., and Prade, H., editors, *Handbook of Fuzzy Sets Series, Approximate Reasoning and Information Systems*, chapter 6. Kluwer.

Düntsch, I. and Gediga, G. (2003). Approximation operators in qualitative data analysis. In de Swart, H., Orlowska, E., Schmidt, G., and Roubens, M., editors, *Theory and Application of Relational Structures as Knowledge Instruments*, volume 2929 of *Lecture Notes in Computer Science*. Springer Verlag, Heidelberg.

Emptoz, H. (1983). *Modèle prétopologique pour la reconnaissance des formes. Applications en neurophysiologie.* Thèse de Doctorat d'Etat, Univ. Claude Bernard, Lyon I, Lyon, France.

Esteva, F., Garcia, P., and Godo, L. (1997). A Modal Account of Similarity-Based Reasoning. *International Journal of Approximate Reasoning*, 16: 235–260.

Everett, C. J. (1944). Closure Operators and Galois Theory in Lattices. *Trans. Amer. Math. Soc.*, 55:514–525.

Fujio, M. and Bloch, I. (2004). Non-Classical Logic via Mathematical Morphology. Technical Report 2004D010, Ecole Nationale Supérieure des Télécommunications.

Galton, A. (2000). Continous Motion in Discrete Space. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *7th International Conference on Principles of Knowledge Representation and Reasoning KR 2000*, pages 26–37, Breckenridge, CO. Morgan Kaufmann, San Francisco, CA.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis*. Springer Verlag.

Giardina, C. R. and Sinha, D. (1989). Image Processing using Pointed Fuzzy Sets. In *SPIE Intelligent Robots and Computer Vision VIII: Algorithms and Techniques*, volume 1192, pages 659–668.

Ginsberg, M. L., Parkes, A. J., and Roy, A. (1998). Supermodels and Robustness. In *Fifteenth National Conference on Artificial Intelligence AAAI'98*, pages 334–339, Madison, Wisconsin.

Goetcherian, V. (1980). From Binary to Grey Tone Image Processing using Fuzzy Logic Concepts. *Pattern Recognition*, 12:7–15.

Hadwiger, H. (1950). Minkowskische Addition und Subtraktion beliebiger Punktmengen und die Theoreme von Erhard Schmidt. *Math. Z.*, 53:210–218.

Heijmans, H. J. A. M. (1991). Theoretical Aspects of Gray-Level Morphology. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13:568–582.

Heijmans, H. J. A. M. (1994). *Morphological Image Operators*. Academic Press, Boston.

Heijmans, H. J. A. M. (1997). Composing Morphological Filters. *IEEE Trans. Image Processing*, 6(5):713–723.

Heijmans, H. J. A. M. and Ronse, C. (1990). The Algebraic Basis of Mathematical Morphology – Part I: Dilations and Erosions. *Computer Vision, Graphics and Image Processing*, 50:245–295.

Hughes, G. E. and Cresswell, M. J. (1968). *An Introduction to Modal Logic*. Methuen, London, UK.

Jeansoulin, R. and Mathieu, C. (1994). Une logique modale des inférences spatiales. *Revue Internationale de Géomatique*, 4:369–384.

Katsuno, H. and Mendelzon, A. O. (1991). Propositional Kowledge Base Revision and Minimal Change. *Artificial Intelligence*, 52:263–294.

Kaufmann, A. and Gupta, M. M. (1988). *Fuzzy Mathematical Models in Engineering and Management Science*. North-Holland, Amsterdam.

Keller, J. M. and Wang, X. (1995). Comparison of Spatial Relation Definitions in Computer Vision. In *ISUMA-NAFIPS'95*, pages 679–684, College Park, MD.

Konieczny, S. and Pino-Pérez, R. (1998). On the Logic of Merging. In *6th International Conference on Principles of Knowledge Representation and Reasoning*, pages 488–498, Trento, Italy.

Krishnapuram, R., Keller, J. M., and Ma, Y. (1993). Quantitative Analysis of Properties and Spatial Relations of Fuzzy Image Regions. *IEEE Transactions on Fuzzy Systems*, 1(3):222–233.

Kutz, O., Sturm, H., Suzuki, N.-Y., Wolter, F., and Zakharyaschev, M. (2002). Axiomatizing Distance Logics. *Journal of Applied Non Classical Logics*, 12(3–4):425–440.

Lafage, C. and Lang, J. (2000). Logical Representation of Preferences for Group Decision Making. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *7th International Conference on Principles of Knowledge Representation and Reasoning KR 2000*, pages 457–468, Breckenridge, CO. Morgan Kaufmann, San Francisco, CA.

Laplante, P. A. and Giardina, C. R. (1991). Fast Dilation and Erosion of Time Varying Grey Valued Images with Uncertainty. In *SPIE Image Algebra and Morphological Image Processing II*, volume 1568, pages 295–302.

Lin, T. Y. (1995). Neighborhood Systems: A Qualitative Theory for Fuzzy and Rough Sets. In *2nd Annual Joint Conference on Information Science*, pages 255–258, Wrightsville Beach, NC.

Lin, T. Y. and Liu, Q. (1994). Rough Approximate Operators: Axiomatic Rough Sets Theory. In Ziarko, W. P., editor, *Rough Sets, Fuzzy Sets and Knowledge Discovery*, pages 256–260. Springer Verlag, London.

Maragos, P. (2005). Lattice Image Processing: A Unification of Morphological and Fuzzy Algebraic Systems. *Journal of Mathematical Imaging and Vision*, 22:333–353.

Maragos, P., Tzouvaras, V., and Stamou, G. (2001). Synthesis and Applications of Lattice Image Operators Based on Fuzzy Norms. In *IEEE Int'l Conference on Image Processing (ICIP-2001)*, pages 521–524, Thessaloniki, Greece.

Matheron, G. (1975). *Random Sets and Integral Geometry*. J. Wiley & Sons, New York.

Matsakis, P. and Wendling, L. (1999). A New Way to Represent the Relative Position between Areal Objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(7):634–642.

Menger, K. (1942). Statistical Metrics. *Proc. National Academy of Sciences USA*, 28:535–537.

Minkowski, H. (1903). Volumen und Oberfläche. *Math. Ann.*, 57:447–495.

Miyajima, K. and Ralescu, A. (1994). Spatial Organization in 2D Segmented Images: Representation and Recognition of Primitive Spatial Relations. *Fuzzy Sets and Systems*, 65:225–236.

Nachtegael, M. and Kerre, E. E. (2000). Classical and Fuzzy Approaches towards Mathematical Morphology. In Kerre, E. E. and Nachtegael, M., editors, *Fuzzy Techniques in Image Processing*, Studies in Fuzziness and Soft Computing, chapter 1, pages 3–57. Physica-Verlag, Springer.

Nachtegael, M., Kerre, E. E., and Raszikowska, A. M. (2000). On Links between Fuzzy Morphology and Fuzzy Rough Sets. In *Information Processing and Management of Uncertainty IPMU 2000*, pages 1381–1388, Madrid.

Nakatsuyama, M. (1993). Fuzzy Mathematical Morphology for Image Processing. In *ANZIIS-93*, pages 75–79, Perth, Western Australia.

Ore, O. (1944). Galois Connexions. *Trans. Amer. Math. Soc.*, 55:493–513.

Orlowska, E. (1993). Rough Set Semantics for Non-Classical Logics. In *International Workshop on Rough Sets, Fuzzy Sets and Knowledge Discovery*, pages 143–148, Banff, Canada.

Pawlak, Z. (1982). Rough Sets. *International Journal of Information and Computer Science*, 11(5):341–356.

Pawlak, Z. (1987). Rough Logic. *Bulletin of the Polish Academy of Sciences*, 35(5-6):253–258.

Perchant, A. and Bloch, I. (2002). Fuzzy Morphisms between Graphs. *Fuzzy Sets and Systems*, 128(2):149–168.

Perchant, A., Boeres, C., Bloch, I., Roux, M., and Ribeiro, C. (1999). Model-based Scene Recognition Using Graph Fuzzy Homomorphism Solved by Genetic Algorithm. In *GbR'99 2nd International Workshop on Graph-Based Representations in Pattern Recognition*, pages 61–70, Castle of Haindorf, Austria.

Pino-Pérez, R. and Uzcátegui, C. (1999). Jumping to Explanations versus jumping to Conclusions. *Artificial Intelligence*, 111:131–169.

Polkowski, L. (1998). Rough Set Approach to Mathematical Morphology: Approximate Compression of Data. In *Information Processing and Management of Uncertainty IPMU'98*, pages 1183–1189, Paris.

Popov, A. T. (1995). Morphological Operations on Fuzzy Sets. In *IEE Image Processing and its Applications*, pages 837–840, Edinburgh, UK.

Randell, D., Cui, Z., and Cohn, A. (1992). A Spatial Logic based on Regions and Connection. In Nebel, B., Rich, C., and Swartout, W., editors, *Principles of Knowledge Representation and Reasoning KR'92*, pages 165–176, San Mateo, CA. Kaufmann.

Renz, J. and Nebel, B. (2001). Efficient Methods for Qualitative Spatial Reasoning. *Journal of Artificial Intelligence Research*.

Ronse, C. (1990). Why Mathematical Morphology Needs Complete Lattices. *Signal Processing*, 21(2):129–154.

Ronse, C. (2003). Flat Morphological Operators on Arbitrary Power Lattices. In Asano, T., Klette, R., and Ronse, C., editors, *LNCS 2616 Geometry, Morphology, and Computational Imaging, 11th International Workshop on Theoretical Foundations of Computer Vision, Dagstuhl Castle, Germany, April 7–12, 2002, Revised Papers*, pages 1–21. Springer.

Ronse, C. and Heijmans, H.J.A.M. (1991). The Algebraic Basis of Mathematical Morphology – Part II: Openings and Closings. *Computer Vision, Graphics and Image Processing: Image Understanding*, 54:74–97.

Ronse, C. and Tajine, M. (2004). Morphological Sampling of Closed Sets. *Image Analysis & Stereology*, 23:89–109.

Rosenfeld, A. (1979). Fuzzy Digital Topology. *Information and Control*, 40: 76–87.

Rosenfeld, A. (1984). The Fuzzy Geometry of Image Subsets. *Pattern Recognition Letters*, 2:311–317.

Rosenfeld, A. (1985). Distances between Fuzzy Sets. *Pattern Recognition Letters*, 3:229–233.

Rosenfeld, A. and Kak, A. C. (1976). *Digital Picture Processing*. Academic Press, New-York.

Schonfeld, D. and Goutsias, J. (1991). Optimal Morphological Pattern Restoration from Noisy Binary Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13:14–29.

Schweizer, B. and Sklar, A. (1963). Associative Functions and Abstract Semigroups. *Publ. Math. Debrecen*, 10:69–81.

Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press, London.

Serra, J., editor (1988). *Image Analysis and Mathematical Morphology, Vol. 2: Theoretical Advances*. Academic Press, London.

Sinha, D. and Dougherty, E. (1992). Fuzzy Mathematical Morphology. *Journal of Visual Communication and Image Representation*, 3(3):286–302.

Sinha, D. and Dougherty, E. R. (1993). Fuzzification of Set Inclusion: Theory and Applications. *Fuzzy Sets and Systems*, 55:15–42.

Sinha, D., Sinha, P., Dougherty, E. R., and Batman, S. (1997). Design and Analysis of Fuzzy Morphological Algorithms for Image Processing. *IEEE Trans. on Fuzzy Systems*, 5(4):570–584.

Soille, P. (2003). *Morphological Image Analysis: Principles and Applications (2nd edition)*. Springer Verlag.

Varzi, A. (1996). Parts, Wholes, and Part-Whole Relations: The Prospects of Mereotopology. *Data and Knowledge Engineering*, 20(3):259–286.

Yager, R. R. (1991). Connectives and Quantifiers in Fuzzy Sets. *Fuzzy Sets and Systems*, 40:39–75.

Yao, T. Y. (1998). Contructive and Algebraic Methods of the Theory of Rough Sets. *Journal of Information Science*, 109:21–47.

Yao, Y. Y. and Lin, T. Y. (1996). Generalization of Rough Sets using Modal Logics. *Intelligent Automation and Soft Computing*, 2(2):103–120.

Zadeh, L. A. (1975). The Concept of a Linguistic Variable and its Application to Approximate Reasoning. *Information Sciences*, 8:199–249.

Zadeh, L. A. (1979). Fuzzy Sets and Information Granularity. In Gupta, M., Ragade, R., and Yager, R., editors, *Advances in Fuzzy Set Theory and Applications*, pages 3–18. North-Holland, Amsterdam.

Chapter 15

SPATIAL REASONING AND ONTOLOGY: PARTS, WHOLES, AND LOCATIONS

Achille C. Varzi
Columbia University

Second Reader

Antony Galton
University of Exeter

Spatial reasoning is no abstract business. It is, to a great extent, reasoning about entities *located in* space, and such entities have spatial *structure*. If the table is in the kitchen, then it follows that the table top is in the kitchen, too, and it follows because the top is *part of* the table. If the concert took place at the stadium, then it didn't take place in the theater, for concerts are spatially *continuous*. Even when we reason about empty places, we typically do so with an eye to the anatomy of their potential tenants: space as such is perceptually remote and we can hardly understand its structure without imagining what *could* fill the void.

This general feature of our spatial competence might suggest a deep metaphysical truth, to the effect that concrete entities such as objects and events are fundamentally prior to, and independent of, their spatial receptacles. It might even suggest that space itself is just a fiction, a picture of some kind: really there are only objects and events spatially related to one another in various ways. Such was, for instance, the gist of Leibniz's stern *relationism* against Newton's *substantivalism*, in spite of the major role the idea of space plays in the sciences. At the same time, one might argue that our understanding of the spatial structure of objects and events, including their spatial relationhips, depends significantly on our understanding of the structure of space *per se*: that the spatial features we attribute to objects and events are somehow inherited from those of the spatial regions they occupy. Thus, for example, we are inclined to say that ordinary objects have parts insofar as their spatial regions have

M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.), Handbook of Spatial Logics, 945–1038.
© 2007 *Springer.*

parts. We may be inclined to say that the top is part of the table because of its salience and functional role, but we may just as easily talk about the top half of a sphere, or its inner parts, in spite of their lacking any cognitive or functional salience: we identify (and reify) such parts in terms of the parts of the region the sphere occupies.

This tension (not to say this ambiguity) between concrete object-oriented thinking and abstract space-oriented thinking is responsible for many of the philosophical issues that lie behind any formal theory of spatial reasoning. On the one hand, it is natural (if not necessary) to supplement the theory with an explicit account of what kinds of thing may enter into its scope, an account of the sorts of entity that can be located in space—in short, an account of what may be collected under the rubric of 'spatial entities'. On the other hand, we also want the theory to be independent of any specific ontological biases we might have. Whatever spatial entities we are inclined to build into the basic furniture of the world—subatomic particles, middle-size objects of the garden variety, large scattered entities such as crowds, forests, archipelagos, galaxies—our reasoning about their spatial properties and relations appears to be governed by the same general principles, and it is natural to think that such principles must reflect our understanding of the structural features of the spatial environment in which such entities are located. In short, although a theory of spatial reasoning may be viewed as an example of applied logic, the logic of certain spatial concepts and relations, it may also be regarded as an example of a formal theory whose principles do not necessarily depend on the intended domain of application except to the extent that the domain must include entities that properly qualify as *spatial* entities of some sort.

The purpose of this chapter is to take a closer look at these delicate matters. Rather than doing this in general, however, we shall look at how the subtle interplay between purely spatial intuitions and intuitions about concrete spatial entities shows up in the construction of a formal theory. More specifically, we shall consider three sorts of theory, each of which occupies a prominent position in recent literature: (1) mereology, or the theory of parthood relations; (2) topology, broadly understood as a theory of qualitative spatial relations such as continuity and contiguity; and (3) the theory of location proper, which deals explicitly with the relationship between an entity and the spatial region it occupies. Arguably, such theories may be viewed as jointly contributing to an overall appraisal of our spatial competence, and over the last few years there has been considerable progress in each direction. At this point there is some need for a philosophical pause, and our purpose in this chapter is to go some way in the direction of a systematic assessment.

1. Philosophical issues in mereology

Let us begin with mereology. This is often defined as the theory of the part-whole relation, but such a definition is misleading. It suggests that mereology has something to say about both parts and wholes, which is not true. As we shall see in Sec. 2, the notion of a whole goes beyond the conceptual resources of mereology and calls for topological concepts and principles of various sorts. By itself, mereology is best understood as the theory of the parthood relation, regardless of whether the second term of the relation may be said to qualify as a *whole* entity. Thus, for instance, it is a mereological fact that my hand is part of my arm just as it is a mereological fact that it is part of my body, although it may plausibly be argued that only my body qualifies as a whole (maximally connected) object; my arm doesn't.

It is also worth pointing out that mereology has a long pedigree, which makes it a central chapter, not only of formal theories concerned with spatial reasoning, but of any theory in the realm of formal logic and ontology broadly understood. Its roots can be traced back to the early days of philosophy, beginning with the Pre-Socratic atomists and continuing throughout the writings of Plato (especially the *Parmenides* and the *Thaetetus*), Aristotle (the *Metaphysics*, but also the *Physics*, the *Topics*, and *De partibus animalium*), and Boethius (*In Ciceronis Topica*). Mereology occupies a prominent role also in the writings of medieval ontologists and scholastic philosophers such as Peter Abelard, Thomas Aquinas, Raymond Lull, and Albert of Saxony, as well as in Jungius's *Logica Hamburgensis* (1638), Leibniz's *Dissertatio de arte combinatoria* (1666) and *Monadology* (1714), and Kant's early writings (especially the *Monadologia physica* of 1756). As a formal theory of the parthood relation, however, mereology made its way into our times mainly through the work of Franz Brentano and of his pupils, especially Husserl's third *Logical Investigation* (1901). The latter may rightly be considered the first attempt at a rigorous formulation of the theory, though in a format that makes it difficult to disentangle the analysis of mereological concepts from that of other formal notions (such as the relation of ontological dependence). It is not until Leśniewski's *Foundations of a General Theory of Manifolds* (1916) that a pure theory of parthood was given an exact formulation. And because Leśniewski's work was largely inaccessible to non-speakers of Polish, it is only with the publication of Leonard and Goodman's *The Calculus of Individuals* (1940) that mereology has become a chapter of central interest for modern ontologists and logicians. Indeed, although Leśniewski's and Leonard and Goodman's theories came in different logical guises, they are sufficiently similar to be recognized as a common basis for most subsequent developments. The question that interests us here is how such developments—and the variety of motivations that lie behind

them—reflect and affect our understanding of mereology as a formal theory of spatial reasoning.[1]

1.1 'Part' and parthood

To this end, the first thing to observe is that the word 'part' has many different meanings in ordinary language, not all of which correspond to the same relation. In a way, it can be used to indicate any portion of a given entity, regardless of whether the portion itself is attached to the remainder, as in (1), or undetached, as in (2); cognitively salient, as in (1)–(2), or arbitrarily demarcated, as in (3); self-connected, as in (1)–(3), or disconnected, as in (4); homogeneous, as in (1)–(4), or gerrymandered, as in (5); material, as in (1)–(5), or immaterial, as in (6); extended, as in (1)–(6), or unextended, as in (7); spatial, as in (1)–(7), or temporal, as in (8); and so on.

(1) The handle is part of the mug.
(2) This cap is part of your pen.
(3) The left half is your part of the cake.
(4) The cutlery is part of the tableware.
(5) This stuff is only part of what I bought.
(6) That area is part of the living room.
(7) The outermost points are part of the perimeter.
(8) The first act was the best part of the play.

All of these cases illustrate the notion of parthood that forms the focus of mereology. Often, however, the English word 'part' is used in a restricted sense. For instance, it may be used to designate only the cognitively salient relation illustrated in (1) and (2). In this sense, the parts of an object x are just its 'components', i.e., those parts that are available as individual units regardless of their interaction with the other parts of x. (A component is *a part* of an object, rather than just *part* of it; see e.g. Tversky 1989). Clearly, the properties of such restricted relations may not coincide with those of parthood broadly understood, so the principles of mereology should not be expected to carry over automatically.

Also, the word 'part' is sometimes used in a broader sense, for instance to designate the relation of material constitution, as in (9), or the relation of mixture composition, as in (10), or even a relation of conceptual inclusion, as in (11):

(9) The clay is part of the statue.

[1] For a historical survey of mereology, see Henry (1991), Burkhardt and Dufour (1991), and Simons (1991c). For systematic comparisons, see Simons (1987) and Ridder (2002).

(10) Gin is part of martini.
(11) Writing detailed comments is part of being a good referee.

The mereological status of these relations, however, is controversial. The constitution relation exemplified in (9) was included by Aristotle in his threefold taxonomy (*Metaphysics*, Δ, 1023b), but many contemporary authors would rather construe it as a *sui generis*, non-mereological relation (see Rea 1997 and references therein).[2] Similarly, the ingredient-mixture relationship exemplified in (10) is subject to controversy, as the ingredients may involve significant structural connections besides spatial proximity and may therefore fail to retain important characteristics they have in isolation (see Sharvy 1983). As for statements such as (11), it may simply be contended that the term 'part' appears only in the surface grammar and disappears at the level of logical form, for instance if (11) is paraphrased as 'A good referee is one who writes detailed comments'. For more examples and tentative taxonomies, see Winston *et al.* (1987), Gerstl and Pribbenow (1995), and Iris *et al.* (1988).

Finally, it is worth stating explicitly that mereology is typically construed as a piece of formal ontology, i.e., a theory of certain formal properties and relations that are exemplified across a wide range of domains, *whatever the nature of the entities in question*. Thus, although both Leśniewski's and Leonard and Goodman's original theories betray a nominalistic stand, reflecting a conception of mereology as a parsimonious alternative to set theory,[3] most contemporary formulations assume no ontological restriction on the field of 'part'. The relata can be individual entities as in (1)–(8), but also abstract entities such as propositions, sets, types, or properties, as in:

(12) That premise is part of my argument.
(13) The domain of quantification is part of the model.
(14) The colon is part of the title.
(15) Humanity is part of personhood.

(The example in (11) may perhaps be read as expressing a mereological relation between properties, too.) This 'ontological innocence' of mereology plays of course an important role in the appraisal of what principles should hold unrestrictedly: greater generality means fewer axioms, and here the tension between the tasks of an applied logic and those of a purely formal theory shows up most vividly. In the following we focus primarily on the spatially salient

[2]Actually, if the statue is identified with the lump of clay, as some would argue (e.g. Noonan 1993 vs. Johnston 1993), and if identity is treated as a limit case of (improper) parthood, as we shall indeed suppose, then the relation of material constitution is a mereological relation. This, however, is the subject of controversy and we shall come back to it in due time.

[3]To be sure, the original calculus of individuals had variables for classes; the class-free version is due to Goodman (1951). On the link between mereology and nominalism, see Eberle (1970).

uses of 'part', but it is important to keep this tension in mind when it comes to assessing the philosophical underpinnings of the most controversial tenets of mereology.

1.2 Basic principles

With these provisos, let us proceed to unpack the theory. We may ideally distinguish two sorts of mereological principles. On the one hand, there are principles that may be thought of as purely 'lexical' axioms fixing the intended meaning of the relational predicate 'part'. On the other, there are principles that go beyond the obvious and aim at greater sophistication and descriptive power. Exactly where the boundary should be drawn, however, is by itself a matter of controversy.

1.2.1 Parthood as a partial ordering. The obvious is this: regardless of how one feels about matters of ontology, if 'part' stands for the general relation exemplified by all of (1)–(8) above, then it stands for a partial ordering—a reflexive, transitive, antisymmetric relation:

(16) Everything is part of itself.
(17) Any part of any part of a thing is itself part of that thing.
(18) Two distinct things cannot be part of each other.

As it turns out, virtually every theory put forward in the literature accepts (16)–(18), though it is worth mentioning some misgivings that have occasionally been raised.

Concerning reflexivity, one might observe that many legitimate senses of 'part' do not countenance saying that a whole is a part of itself. For instance, Rescher (1955, p. 10) cited the biologists' use of 'part' for the functional subunits of an organism as a case in point. This is of little import, though. Taking reflexivity as constitutive of the meaning of 'part' amounts to regarding identity as a limit (improper) case of parthood. A stronger relation, whereby nothing counts as part of itself, can obviously be defined in terms of the weaker one, hence there is no loss of generality (see Sec. 1.2.2). Vice versa, one could frame a mereological theory by taking proper parthood as a primitive instead. This is merely a question of choosing a suitable primitive.

The transitivity principle, (17), is more controversial. Several authors have observed that many legitimate senses of 'part' are non-transitive. Examples would include: (i) a biological subunit of a cell is not a part of the organ of which that cell is a part; (ii) a handle can be part of a door and the door of a house, though a handle is never part of a house; (iii) my finger is part of me and I am part of the team, yet my finger is not part of the team. (See again Rescher 1955, Cruse 1979, and Winston *et al.* 1987, respectively; for other examples see Iris *et al.* 1988, Moltman 1997, and Johansson 2004 *inter alia*). Arguably,

however, such misgivings stem again from the ambiguity of 'part'. What counts as a biological subunit of a cell may not count as a subunit, i.e., a *distinguished* part of the organ, but that is not to say that it is not part of the organ at all. Similarly, if there is a sense of 'part' in which a handle is not part of the house to which it belongs, or my finger not part of my team, it is a restricted sense: the handle is not a *functional* part of the house (a 'component'), though it is a functional part of the door and the door a functional part of the house; my finger is not *directly* part of the team, though it is directly part of me and I am directly part of the team. It is obvious that if the interpretation of 'part' is narrowed by additional conditions (e.g., by requiring that parts make a functional or direct contribution to the whole), then transitivity may fail. In general, if x is a ϕ-part of y and y is a ϕ-part of z, x need not be a ϕ-part of z: the predicate modifier 'ϕ' may not distribute over parthood. But that shows the non-transitivity of 'ϕ-part', not of 'part'. And within a sufficiently general framework this can easily be expressed with the help of explicit predicate modifiers (Varzi 2005). In any event, it seems clear that *spatial* parthood is transitive: whether we construe this as a restricted notion or identify it with the general notion of parthood, (17) holds.

Finally, concerning the antisymmetry postulate (18), two sorts of worry are worth mentioning. On the one hand, some authors maintain that the relationship between an object and the stuff it is made of provides a perfectly ordinary example of symmetric parthood: according to Thomson (1998), for example, a statue and the clay it is made of are part of each other, yet distinct. This is highly controversial and there is a large philosophical literature devoted on this topic (see e.g. the papers in Rea 1997). For the moment, let us simply observe that the example trades once again on the ambiguity of 'part'. We have already mentioned that material constitution is best regarded as a *sui generis*, non-mereological relation. Whether this relation may obtain between two spatially coincident objects is an interesting question, but we should postpone its discussion to where it belongs: the theory of spatial location (Sec. 3.3). On the other hand, one may wonder about the possibility of *unordinary* cases of symmetric parthood relationships. Sanford (1993, p. 222) refers to Borges's *Aleph* as a case in point: 'I saw the earth in the Aleph and in the earth the Aleph once more and the earth in the Aleph ...'. In this case, a plausible reply is simply that fiction delivers no guidance to conceptual investigations: conceivability may well be a guide to possibility, but literary fantasy is by itself no evidence of genuine conceivability (van Inwagen 1993, p. 229). Still, one may observe that the possibility of mereological loops is not pure fantasy. In view of certain developments in non-well-founded set theory (Aczel 1988), one might indeed suggest building mereology on the basis of a notion of parthood that may violate (18). This is particularly significant insofar as set theory itself may be reformulated in mereological terms—a possibility that is explored in the works of

Bunt (1985) and especially Lewis (1991). At present, however, no systematic study of non-well-founded mereology has been put forward in the literature. Moreover, we are interested here in mereology as a tool for spatial reasoning, and in this regard the possibility of symmetric loops does indeed appear to be pure fantasy. In the following we shall therefore confine ourselves to theories that accept the antisymmetry postulate along with reflexivity and transitivity: parthood is a partial ordering.

1.2.2 Other mereological concepts. It is convenient at this point to introduce some degree of formalization. Let us use 'P' for the binary predicate constant '...is part of ...'. Taking the underlying logic to be a standard predicate calculus with identity, the above minimal requisites on parthood may then be regarded as forming a first-order theory characterized by the following proper axioms for 'P':

(P.1)	$\mathsf{P}xx$	Reflexivity
(P.2)	$\mathsf{P}xy \wedge \mathsf{P}yz \rightarrow \mathsf{P}xz$	Transitivity
(P.3)	$\mathsf{P}xy \wedge \mathsf{P}yx \rightarrow x = y$	Antisymmetry

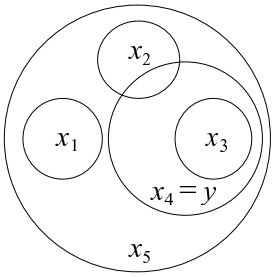
(Here and in the following we simplify notation by dropping all initial universal quantifiers. Unless otherwise specified, all formulas are to be understood as universally closed.) A number of additional mereological predicates can then be introduced by definition. For example:

(19)	$\mathsf{EQ}xy =_{df} \mathsf{P}xy \wedge \mathsf{P}yx$	Equality
(20)	$\mathsf{PP}xy =_{df} \mathsf{P}xy \wedge \neg\mathsf{P}yx$	Proper Parthood
(21)	$\mathsf{PE}xy =_{df} \neg\mathsf{P}xy \wedge \mathsf{P}yx$	Proper Extension
(22)	$\mathsf{O}xy =_{df} \exists z(\mathsf{P}zx \wedge \mathsf{P}zy)$	Overlap
(23)	$\mathsf{U}xy =_{df} \exists z(\mathsf{P}xz \wedge \mathsf{P}yz)$	Underlap

An intuitive model for these relations, with 'P' interpreted as spatial inclusion, is given in the diagram of Fig. 15.1.

Note that '$\mathsf{U}xy$' is bound to hold if we assume the existence of a 'universal entity' of which everything is part. Conversely, '$\mathsf{O}xy$' would always hold if we assumed the existence of a 'null entity' that is part of everything. In the domain of spatial entities, the latter assumption is of course implausible (Geach 1949).[4] The former assumption may be challenged, too (Simons 2003, Varzi 2006), but it seems reasonable, if not obvious, in case the only spatial entities countenanced by the theory are regions of space. We shall come back to these issues in Sec. 1.4.

[4]In other contexts one may feel differently: see Martin (1965) and Bunt (1985) for theories with a null individual, and Bunge (1966) for a theory with *several* null individuals.



	U$x_n y$	O$x_n y$	PP$x_n y$	EQ$x_n y$	PE$x_n y$
$n = 1$	+	–	–	–	–
$n = 2$	+	+	–	–	–
$n = 3$	+	+	+	–	–
$n = 4$	+	+	–	+	–
$n = 5$	+	+	–	–	+

Figure 15.1. Basic mereological relations. (Shaded cells indicate parthood.)

Note also that the definitions imply (by pure logic) that EQ, O, and U are reflexive and symmetric; in addition, EQ is also transitive—an equivalence relation. By contrast, PP and PE are irreflexive and asymmetric, and it follows from (P.2) that both are transitive. Since the following biconditional is also a straightforward consequence of the axioms:

(24) $\mathsf{P}xy \leftrightarrow (\mathsf{PP}xy \vee x = y)$

it should now be obvious that one could in fact use proper parthood as an alternative starting point for the development of mereology, using the right-hand side of (24) as a definiens for 'P'. This is, for instance, the option followed in Simons (1987), where the partial ordering axioms for 'P' are replaced by the strict ordering axioms for 'PP':

(25) $\neg\mathsf{PP}xx$
(26) $\mathsf{PP}xy \wedge \mathsf{PP}yz \rightarrow \mathsf{PP}xz$
(27) $\mathsf{PP}xy \rightarrow \neg\mathsf{PP}yx$

Ditto for 'EP', which was in fact the primitive relation in Whitehead's (1919) semi-formal treatment of the mereology of events. Other options may be considered, too. For example, Goodman (1951) used 'O' as a primitive and Leonard and Goodman (1940) used its opposite:

(28) $\mathsf{D}xy =_{df} \neg\mathsf{O}xy$ Disjointness

However, the relations corresponding to such predicates are weaker than PP and PE and no biconditional is provable from (P.1)–(P.3) that would yield a corresponding definiens of 'P' (though one could of course define 'P' in terms of 'O' or 'D' in the presence of further axioms; see below *ad* (45)). Thus, other things being equal, 'P', 'PP', and 'PE' appear to be the only reasonable options. Here we shall stick to 'P'.

Finally, note that Identity could itself be introduced by definition, due to the following corollary of the antisymmetry postulate (P.3):

(29) $x = y \leftrightarrow \mathsf{EQ}xy$

Accordingly, the theory could be formulated in a pure first-order language by assuming (P.1) and (P.2) and replacing (P.3) with the following variant of the standard axiom schema for '=' (where ϕ is any formula):

(P.3′) $\mathsf{EQ}xy \rightarrow (\phi x \leftrightarrow \phi y)$ Indiscernibility

One may in fact argue on these grounds that parthood is in some sense conceptually prior to identity (as in Sharvy 1983, p. 234), and since 'EQ' is not definable in terms of 'PP' or 'PE' without resorting to '=', the argument would also provide evidence in favor of 'P' as the most basic primitive. As we shall see in Sec. 1.3.2, however, the link between parthood and identity is philosophically problematic. In order not to compromise our discussion, in the following we shall therefore continue to work with a language with both 'P' and '=' as primitives.

1.3 Decomposition principles

Let M be the theory defined by the three basic principles (P.1)–(P.3). M may be viewed as embodying the common core of any mereological theory. Not just any partial ordering qualifies as a part-whole relation, though, and deciding what further principles should be added to (P.1)–(P.3) is precisely the question a good mereological theory is meant to answer. It is here that philosophical issues begin to arise.

Generally speaking, such refinements may be divided into two main groups. On the one hand, one may extend M by means of *decomposition principles* that take us from a whole to its proper parts. For example, one may consider the idea that whenever something has a proper part, it has more than one—i.e., that there is always some *mereological difference* (a remainder) between a whole and its proper parts. This need not be true in every model for M: a world with only two items, only one of which is part of the other, would be a counterexample, though not one that could be illustrated with the sort of geometric diagram used in Fig. 15.1. On the other hand, one may extend M by means of *composition principles* that go in the opposite direction—from the parts to the whole. For example, one may consider the idea that whenever there are some things there exists a whole that consists exactly of those things—i.e., that there is always a *mereological sum* (or fusion) of two or more parts. Again, this need not be true in a model for M, and it is a matter of controversy whether the idea should hold unrestrictedly.

1.3.1 Parts and remainders. Let us begin with the first sort of extension. And let us start by taking a closer look at the intuition according to which a whole cannot be decomposed into a single proper part. There are various ways

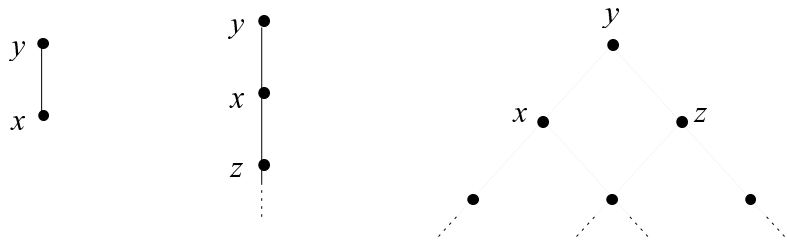


Figure 15.2. Three unsupplemented models. (Parthood relationships are represented by connecting lines going uphill.)

in which one can try to capture this basic intuition. Consider the following (from Simons 1987, pp. 26–28):

(P.4$_a$)	$\mathsf{PP}xy \rightarrow \exists z(\mathsf{PP}zy \wedge \neg z = x)$	Weak Company
(P.4$_b$)	$\mathsf{PP}xy \rightarrow \exists z(\mathsf{PP}zy \wedge \neg \mathsf{P}zx)$	Strong Company
(P.4)	$\mathsf{PP}xy \rightarrow \exists z(\mathsf{P}zy \wedge \neg \mathsf{O}zx)$	Supplementation

The first principle, (P.4$_a$), is a literal rendering of the idea in question: every proper part must be accompanied by another. However, there is an obvious sense in which (P.4$_a$) only captures the letter of the idea, not the spirit: it rules out the unintended model mentioned above (see Fig. 15.2, left) but not, for example, an implausible model with an infinitely descending chain in which the additional proper parts do not leave any remainder (Fig. 15.2, center).

The second principle, (P.4$_b$), is stronger: it rules out both models as unacceptable. However, (P.4$_b$) is still too weak to capture the intended idea. For example, it is satisfied by a model in which a whole can be decomposed into several proper parts all of which overlap one another (Fig. 15.2, right), and it is may be argued that such models do not do justice to the meaning of 'proper part': after all, the idea is that the removal of a proper part should leave a remainder, but it is by no means clear what would be left of z once x (along with its parts) is removed.

It is only the third principle, (P.4), that appears to provide a full formulation of the idea that nothing can have a single proper part. According to this principle, every proper part must be 'supplemented' by another, *disjoint* part, and it is this last qualification that captures the notion of a remainder.

Should (P.4) be incorporated into M as a further fundamental principle on the meaning of 'part'? Most authors (beginning with Simons himself) would say so. Yet here there is room for disagreement. In fact, it is not difficult to conceive of mereological scenarios that violate not only (P.4), but also (P.4$_b$) and even (P.4$_a$). For example, in Brentano's (1933) theory of accidents, a soul is a proper part of a thinking soul even though there is nothing to make up for the difference (see Chisholm 1978; Baumgartner and Simons 1994). Similarly, in Fine's

(1982) theory of *qua* objects, every basic object (John) qualifies as the only proper part of its incarnations (John *qua* philosopher, John *qua* husband, etc.). Now, such putative counterexamples are controversial and, more importantly for our present concerns, they appear to be of little significance if mereology is to be thought of as a theory of space. The spatial relations illustrated by our initial examples (1)–(7) all seem to satisfy (P.4) and, *a fortiori*, (P.4$_a$) and (P.4$_b$). Nonetheless there are counterexamples also in the realm of truly spatial mereologies. The best illustration comes from Whitehead's (1929) theory of extensive connection: on this theory, a topologically closed region includes its open interior as a proper part in spite of there being no boundary elements to distinguish them—the domain only consists of extended regions. Whether the omission of boundary elements such as points, lines, and surfaces is a reasonable thing to do when it comes to the task of modeling our understanding of space, and whether in the absence of such elements the distinction between open and closed regions is still legitimate, are questions that every theory of space must of course address. In Sec. 2.4 we shall see that answering in the affirmative involves serious philosophical and technical complications. But we shall also see that several theories are available to do the job, including theories that occupy a prominent role in the current literature on qualitative spatial reasoning. One may rely on the intuitive appeal of (P.4) to discard such theories as implausible, but one may as well turn things around and regard the adequacy of such theories as a good reason not to accept (P.4) unrestrictedly. As things stand, it therefore seems appropriate to regard such a principle as providing a minimal but substantive addition to (P.1)–(P.3), one that goes beyond the mere lexical characterization of 'part' provided by M. For future reference, let us label the resulting mereological theory *MM* (for *Minimal Mereology*).

1.3.2 Supplementation, extensionality, identity. There is another way of expressing the supplementation intuition that is worth considering. It corresponds to the following axiom, which differs from (P.4) in the antecedent:

(P.5) $\neg \mathsf{P}yx \rightarrow \exists z(\mathsf{P}zy \wedge \neg \mathsf{O}zx)$ Strong Supplementation

Intuitively, this says that if an object *fails* to include another among its parts, then there must be a remainder. It is easily seen that (P.5) implies (P.4), so any theory rejecting at least (P.4) will *a fortiori* reject (P.5). (For instance, on Whitehead's boundary-free theory of extensive connection, a closed region is not part of its interior even though each part or the former overlaps the latter.) However, the converse does not hold. The diagram in Fig. 15.3 illustrates a model in which (P.4) is true, since each proper part counts as a supplement of the other; yet (P.5) is false.

The theory obtained by adding (P.5) to (P.1)–(P.3) is thus a proper extension of MM. Let us label this stronger theory EM, for *Extensional Mereology*,

Figure 15.3. A weakly supplemented model.

the attribute 'extensional' being justified precisely by the exclusion of counter-models that, like the one just mentioned, contain distinct objects with the same proper parts. In fact, the following is a theorem of EM:

(30) $\exists z \mathsf{PP} zx \rightarrow (\forall z(\mathsf{PP} zx \rightarrow \mathsf{PP} zy) \rightarrow \mathsf{P} xy)$

from which it follows that no composite objects with the same proper parts can be distinct:

(31) $(\exists z \mathsf{PP} zx \vee \exists z \mathsf{PP} zy) \rightarrow (x = y \leftrightarrow \forall z(\mathsf{PP} zx \leftrightarrow \mathsf{PP} zy))$

(The analogue for 'P' is, of course, already provable in M, since parthood is reflexive and antisymmetric.) Thus, EM is truly an extensional theory incorporating the view that an object is exhaustively defined by its constituent parts. This goes far beyond the intuition that lies behind the weak supplementation principle (P.4). Does it go too far?

On the face of it, it is not difficult to envisage scenarios that would correspond to the diagram in Fig. 15.3. For example, we can obtain a counterexample to (P.5) by identifying x and y with the sets $\{\{z_1\}, \{z_1, z_2\}\}$ and $\{\{z_2\}, \{z_1, z_2\}\}$ (i.e., with the ordered pairs $\langle z_1, z_2\rangle$ and $\langle z_2, z_1\rangle$, respectively), interpreting 'P' as the ancestral of the improper membership relation (i.e., of the union of $\in$ and $=$). But sets are abstract entities; can we also envisage similar scenarios in the spatial domain?

Here is a case where the answer may differ crucially depending on whether we are interested in modeling a domain of concrete spatial entities or just the domain of the regions of space that they occupy. In the latter case there is little room for controversy: spatial regions are extensional, if anything is, unless of course we favor a Whitheadian conception of space. In the former case, however, the answer is controversial. There are two sorts of objection worth considering. On the one hand, it is sometimes argued that sameness of parts is not *sufficient* for identity, as some entities may differ exclusively with respect to the arrangement of their parts. For example, it is sometimes argued that: (i) two words can be made up of the same letters, as with 'fallout' and 'outfall'; (ii) the same flowers can compose a nice bunch or a scattered bundle, depending on the arrangements of the individual flowers; (iii) a cat can survive the annihilation of its tail, but the amount of feline tissue consisting of the cat's tail and the rest of the cat's body cannot survive the annihilation of the tail, hence they have different

properties and must be distinct by Leibniz's law in spite of their sharing exactly the same ultimate mereological constituents. (See Hempel 1953: 110, Eberle 1970: §2.10, and Wiggins 1968, respectively; variants of (iii) may also be found in Doepke 1982, Lowe 1989, Johnston 1992, and Baker 1999, *inter alia*.) On the other hand, it is sometimes argued that sameness of parts is not *necessary* for identity, as some entities may survive mereological change. If a cat survives the annihilation of its tail, then the tailed cat (before the accident) and the tailless cat (after the accident) are one and the same in spite of their having different proper parts (Wiggins 1980). If any of these arguments is accepted, then clearly (31) is too strong a principle to be imposed on the parthood relation. And since (31) follows from (P.5), it might be concluded that EM is on the wrong track.

Let us look at these objections separately. Concerning the necessity of mereological extensionality, i.e., the left-to-right conditional in the consequent of (31):

(32) $x = y \rightarrow \forall z(\mathsf{PP}zx \leftrightarrow \mathsf{PP}zy)$

it is perhaps enough to remark that the difficulty is not peculiar to extensional mereology. The objection proceeds from the consideration that ordinary entities such as cats and other living organisms (and possibly other entities as well, such as cars and houses) survive all sorts of gradual mereological changes. Yet the same can be said of other types of change as well: bananas ripen, houses deteriorate, people sleep at night and eat at lunch. How can we say that they are the same things, if they are not quite the same? Indeed, (32) is just an instance of the identity axiom

(ID) $x = y \rightarrow (\phi x \leftrightarrow \phi y)$

and it is well known that this axiom calls for revisions when '=' is given a diachronic reading. Arguably, any such revisions will affect the case at issue as well, and in this sense the above-mentioned objection to (32) can be disregarded. For example, if the basic parthood predicate were reinterpreted as a time-indexed relation (Thomson 1983), then the problem would disappear as the tensed version of (P.5) would only warrant the following variant of (32):

(32′) $x = y \rightarrow \forall t \forall z(\mathsf{PP}_t zx \leftrightarrow \mathsf{PP}_t zy)$

Similarly, the problem would disappear if the variables in (32) were taken to range over four-dimensional entities whose parts may extend in time as well as in space (Heller 1984, Sider 2001), or if identity itself were construed as a contingent relation that may hold at some times but not others (Gallois 1998). Such revisions may be regarded as an indicator of the limited ontological neutrality of extensional mereology. But their independent motivation also bears witness to the fact that controversies about the necessity of extensionality

stem from larger and more fundamental philosophical conundrums and cannot be assessed by appealing to our intuitions about the meaning of 'part'.

The worry about the sufficiency of mereological extensionality, i.e., the right-to-left conditional in the consequent of (31):

(33) $\forall z(\mathsf{PP}zx \leftrightarrow \mathsf{PP}zy) \rightarrow x = y$

is more to the point. However, there are various ways of resisting such counterexamples as (i)–(iii) on behalf of EM. Consider (i)—two words made up of the same letters. Insofar as we are dealing with truly spatial entities, this is best described as a case of different word *tokens* made up of distinct tokens of the same letter *types*. There is, accordingly, no genuine violation of (33) in the opposition between 'fallout' and 'outfall' (for instance), hence no reason to reject (P.5) on these grounds. (Besides, even with respect to abstract types, it could be pointed out that 'fallout' and 'outfall' do not share *all* their proper parts: the string 'lo', for instance, is only included in the first word.) What if one of the two word tokens is obtained from the other by rearranging the same letter *tokens*? In that case the reply misfires, but so does the ojection: the issue becomes once again one of diachronic non-identity, with all that it entails (Lewis 1991, pp. 78ff).

Case (ii)—the flowers—is not significantly different. The same, concrete flowers cannot compose a nice bunch and a scattered bundle *at the same time*. Case (iii), however, is more delicate. There is a strong intuition that a cat really is something over and above the amount of feline tissue consisting of its tail and the rest of its body—that they have different survival conditions and, hence, different properties—so it may be thought that here we have a genuine counterexample to mereological extensionality. On behalf of EM, it should nonetheless be noted that the appeal to Leibniz's law in this context is debatable. Let 'Tibbles' name our cat and 'Tail' its tail, and let us grant the truth of

(34) Tibbles can survive the annihilation of Tail.

There is, indeed, an intuitive sense in which the following is also true:

(35) The amount of feline tissue consisting of Tail and the rest of Tibbles's body cannot survive the annihilation of Tail.

However, this intuitive sense corresponds to a *de dicto* reading of the modality, where the description in (35) has narrow scope:

(35a) Necessarily, the amount of feline tissue consisting of Tail and the rest of Tibbles's body has Tail as a proper part.

On this reading (35) is hardly negotiable (in fact, logically true). Yet this is irrelevant in the present context, for (35a) does not amount to an ascription of a

modal property and cannot be used in connection with Leibniz's law. (Compare the following fallacious argument: The number of planets might have been even; 9 is necessarily odd; hence the number of planets is not 9.) On the other hand, consider a *de re* reading of (35):

(35b) The amount of feline tissue consisting of Tail and the rest of Tibbles's body necessarily has Tail as a proper part.

On this reading the appeal to Leibniz's law would be legitimate (modulo any concerns about the status of modal properties) and one could rely on the truth of (34) and (35), i.e., (35b), to conclude that Tibbles is distinct from the relevant amount of feline tissue. However, there is no obvious reason why (35) should be regarded as true on this reading. That is, there is no obvious reason to suppose that the amount of feline tissue that in the actual world consists of Tail and the rest of Tibbles's body—*that* amount of feline tissue that is now resting on the carpet—cannot survive the annihilation of Tail. Indeed, it would appear that any reason in favor of this claim *vis-à-vis* the truth of (34) would have to *presuppose* the distinctness of the entities in question, so no appeal to Leibniz's law would be legitimate to *establish* the distinctess on pain of circularity (Varzi 2000). This is not to say that the putative counterexample to (34) is wrong-headed. But it requires genuine metaphysical work to defend it and it makes the rejection of the strong supplementation principle (P.5) a much harder task.

1.3.3 Complementation. There is a way of expressing the supplementation intuition that is even stronger than (P.5). It corresponds to the following thesis, which differs from (P.5) in the consequent:

(P.6) $\neg \mathsf{P}yx \rightarrow \exists z \forall w(\mathsf{P}wz \leftrightarrow (\mathsf{P}wy \wedge \neg \mathsf{O}wx))$ Complementation

This says that if y is not part of x, there exists something that comprises exactly those parts of y that are disjoint from x—something that we may call the *difference* or relative *complement* between y and x. It is easily checked that this principle implies (P.5). On the other hand, the diagram in Fig. 15.4 shows that the converse does not hold: there are two parts of y disjoint from x, namely z_1 and z_2, but there is nothing that consists exactly of such parts, so we have a model of (P.5) in which (P.6) fails.

Any misgivings about (P.5) may of course be raised against (P.6). But what if we agree with the above arguments in support of (P.5)? Do they also give us reasons to accept the stronger principle (P.6)? The answer is in the negative. Plausible as it may sound, (P.6) has consequences that even an extensionalist may not be willing to accept. For example, Fig. 15.5 depicts a scenario that—it may be argued—corresponds exactly to the model of Fig. 15.4. It may be argued that although x and z_1 jointly constitute a larger part of y (the difference between y and z_2), and similarly for x and z_2 (the difference between y and

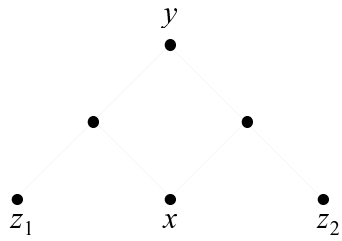


Figure 15.4. A strongly supplemented model violating complementation.

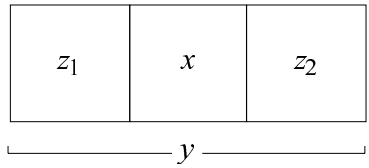


Figure 15.5. Does the difference between y and x exist?

z_1), there is nothing consisting of z_1 and z_2 (the difference between y and x), since these two pieces are disconnected. More generally, it appears that (P.6) would force us to accept the existence of scattered entities, such as the 'sum' of your left and right arms, or the 'sum' of Canada and Mexico, and since Lowe (1953) many authors have objected to this thought *regardless* of how one feels about extensionality. (One philosopher who explicitly agrees to extensionality while distrusting scattered entities is Chisholm 1987.) As it turns out, the extra strength of (P.6) is therefore best appreciated in terms of the sort of mereological aggregates that this principle would entail, aggregates that are composed of two or more parts of a given whole. This suggests that any additional misgivings about (P.6), besides its extensional implications, are truly misgivings about matters of composition. We shall accordingly postpone their discussion to Sec. 1.4, where we shall attend to these matters more fully. For the moment, let us simply say that (P.6) is, on the face of it, not a principle that can be added to M without further argument.

1.3.4 Atomism and other options. One last important family of decomposition principles concerns the question of atomism. Mereologically, an atom (or 'simple') is an entity with no proper parts:

(36) $\mathsf{A}x =_{df} \neg\exists y \mathsf{PP}yx$ Atom

Are there any such entities? And if there are, is everything entirely made up of atoms? Does everything comprise at least some atoms? Or is everything made up of atomless gunk? These are deep and difficult questions, which have been the focus of philosophical investigation since the early days of philosophy

and have been center stage also in many recent disputes in mereology (see, for instance, van Inwagen 1990, Sider 1993, Zimmerman 1996a, and the papers collected in Hudson 2004). Here we shall confine ourselves to a brief examination.

The two main options, to the effect that there are no atoms at all, or that everything is ultimately made up of atoms, correspond to the following postulates, respectively:

(P.7) $\exists y \mathsf{PP} yx$ Atomlessness
(P.8) $\exists y(\mathsf{A}y \wedge \mathsf{P}yx)$ Atomicity

These postulates are mutually incompatible, but taken in isolation they can consistently be added to any mereological theory X considered here. Adding (P.8) yields a corresponding *Atomistic* version, AX; adding (P.7) yields an *Atomless* version, $\bar{A}X$. Since finitude together with the antisymmetry of parthood (P.3) jointly imply that mereological decomposition must eventually come to an end, it is clear that any finite model of M (and *a fortiori* of any extension of M) must be atomistic. Accordingly, an atomless mereology $\bar{A}X$ admits only models of infinite cardinality. (A world containing such wonders as Borges's Aleph, where parthood is not antisymmetric, might by contrast be finite and yet atomless.) An example of such a model, establishing the consistency of the atomless version of most mereological theories considered in the this chapter, is provided by the regular open sets of a Euclidean space, with 'P' interpreted as set-inclusion (Tarski 1935).

Now, one thing to notice is that, independently of their motivations, atomistic mereologies admit of significant simplifications in the axioms. For instance, AEM can be simplified by replacing (P.5) and (P.8) with

(P.5′) $\neg \mathsf{P}yx \rightarrow \exists z(\mathsf{A}z \wedge \mathsf{P}zy \wedge \neg \mathsf{P}zx)$

which in turns implies the following atomistic variant of the extensionality thesis (31):

(37) $x = y \leftrightarrow \forall z(\mathsf{A}z \rightarrow (\mathsf{P}zx \leftrightarrow \mathsf{P}zy))$

Thus, any atomistic extensional mereology is truly 'hyperextensional' in Goodman's (1958) sense: things built up from exactly the same atoms are identical. An interesting question, discussed at some length in the late 1960's (Yoes 1967, Eberle 1968, Schuldenfrei 1969) and taken up more recently by Simons (1987: 44f) and Engel and Yoes (1996), is whether there are atomless analogues of (37). Is there any predicate that can play the role of 'A' in an atomless mereology? Such a predicate would identify the 'base' of the system and would therefore enable mereology to cash out Goodman's hyperextensional intuitions even in the absence of atoms. This question is particularly significant

from a nominalistic perspective, but it also bears on our present concerns. For example, it is a relevant question to ask in connection with the Whiteheadian conception mentioned in Sec. 1.3.1, according to which space contains no parts of lower dimensions such as points or lines (see Forrest 1996, Roeper 1997). In special cases there is no difficulty in providing a positive answer. For example, in the $\bar{A}EM$ model consisting of the open regular subsets of the real line, the open intervals with rational end points form a base in the relevant sense. It is unclear, however, whether a general answer can be given that applies to any sort of domain. If not, then the only option would appear to be an account where the notion of a 'base' is relativized to entities of a given sort. In Simons's terminology, we could say that the ψ-ers form a base for the ϕ-ers if and only if the following variants of (P.5′) and (P.8) and are satisfied:

(P.5*) $\phi x \wedge \phi y \rightarrow (\neg \mathsf{P} yx \rightarrow \exists z(\psi z \wedge \mathsf{P} zy \wedge \neg \mathsf{P} zx))$
(P.8*) $\phi x \rightarrow \exists y(\psi y \wedge \mathsf{P} yx)$

An atomistic mereology would then correspond to the limit case where 'ψ' is identified with 'A' for every choice of 'ϕ'. In an atomless mereology, by contrast, the choice of the base would depend each time on the level of 'granularity' set by the relevant specification of 'ϕ'.

A second important consideration concerns the possibility of theories that lie between the two extreme options afforded by Atomicity and Atomlessness. For instance, it can be held that there are atoms, though not everything need have a complete atomic decomposition, or it can be held that there is atomless gunk, though not everything need be gunky (Zimmerman 1996a). Again, formally this amounts to endorsing a restricted version of either (P.7) or (P.8) in which the variables are suitably restricted so as to range over entities of a certain sort:

(P.7$_\phi$) $\phi x \rightarrow \exists y \mathsf{P}\mathsf{P} yx$
(P.8$_\phi$) $\phi x \rightarrow \exists y(\mathsf{A} y \wedge \mathsf{P} yx)$

At present, no thorough formal investigation of such options has been entertained (but see Masolo and Vieu 1999). Yet the issue is particularly significant from the perspective of a mereological theory aimed at modeling the spatial world, especially if the theory is to countenance concrete spatial entities along with the regions of space that such entities may occupy. It is, after all, a plausible thought that while the question of atomism may be left open with regard to the mereological structure of material objects (pending empirical findings from physics, for example), it must receive a definite answer with regard to the structure of space itself. This would amount to endorsing a version of either (P.7$_\phi$) or (P.8$_\phi$) in which 'ϕ' is understood as a condition that is satisfied exclusively by regions of space. Such a condition, of course, cannot be formulated in the language of a purely mereological theory, but we shall see in Sec. 3

that a suitably enriched theory, in which the relation of location is explicitly articulated, can do the job properly. (Actually, it is hard to conceive of a world in which an atomistic space is inhabited by entities that can be decomposed indefinitely, so in this case it is reasonable to suppose that any theory accepting (P.8$_\phi$) for regions would also endorse the stronger principle (P.8). However, (P.7$_\phi$) would be genuinely independent of (P.7), unless it is assumed that every mereologically atomic entity should also be spatially atomic, i.e., unextended.)

Similar considerations apply to other decomposition principles that may come to mind at this point. For example, one may consider adding a requirement to the effect that 'PP' forms a dense ordering, as already Whitehead (1919) had it:

(P.9) $\mathsf{PP}xy \rightarrow \exists z(\mathsf{PP}xz \wedge \mathsf{PP}zy)$ Density

As a general decomposition principle, (P.9) might be deemed too strong, especially in an atomistic setting. However, it is plausible to suppose that (P.9) should hold at least in the domain of spatial regions, regardless of whether these are construed as atomless gunk or as aggregates of spatial atoms. Evidently much depends on the link one establishes between the mereology of an object and that of its spatial location and this, again, is a question to which we attend more fully in Sec. 3. For the moment, let us simply observe that the sort of philosophical issues that lie behind these options is significantly different from those considered in the previous sections. Whether something can have a single proper part, whether parthood is extensional, or even whether it satisfies the complementation principle (P.6) are issues that depend greatly on our understanding of the parthood relation. They are, in an important sense, conceptual questions. Whether there are mereological atoms, by contrast, or whether mereological decomposition should obey a density principle, are substantive questions that have nothing to do with our understanding of parthood as such. (For more on these questions, and on their general historical background, see Pyle 1995 and Holden 2004.)

1.4 Composition principles

Let us now consider the second way of extending M mentioned at the beginning of Sec. 1.3. Just as we may want to fix the logic of P by means of decomposition principles that take us from a whole to its proper parts, we may look at composition principles that go in the opposite direction—from the parts to the whole. More generally, we may consider the idea that the domain of the theory ought to be closed under mereological operations of various sorts: not only fusions, but also products, differences, and more. Here, again, there is room for several philosophical considerations, some of which are particularly indicative of the tension between space-oriented and object-oriented intuitions.

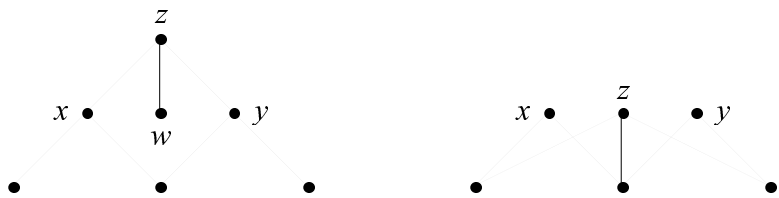


Figure 15.6. A fusion$_a$ that is not a fusion, and a fusion that is not a fusion$_b$.

1.4.1 Bounds and fusions. Conditions on composition are many. Beginning with the weakest, consider the claim that any two suitably related entities have an upper bound, i.e., underlap:

$$(\text{P.10}_\psi) \quad \psi xy \rightarrow \exists z(\mathsf{P}xz \wedge \mathsf{P}yz) \qquad \textsf{Boundedness}$$

Exactly how 'ψ' should be construed is an important question by itself—a version of what van Inwagen (1990) calls the 'special composition question'. Perhaps the most natural choice is to identify ψ with mereological overlap, the rationale being that such a relation establishes an important tie between what may count as two distinct parts of a larger whole. As we shall see momentarily, with ψ so construed, (P.10$_\psi$) is indeed uncontroversial. However, regardless of any specific choice, it is apparent that (P.10$_\psi$) is pretty weak, as it holds trivially in any domain with a universal entity of which everything is part.

A somewhat stronger condition would be to require any pair of suitably related entities to have a *smallest* underlapper—something composed exactly of them and nothing else. This requirement is sometimes stated by saying that such entities must have a mereological 'sum', or 'fusion', though it is not immediately obvious how that should be formulated in the formal language. Consider:

$$(\text{P.11}_{\psi a}) \quad \psi xy \rightarrow \exists z(\mathsf{P}xz \wedge \mathsf{P}yz \wedge \forall w(\mathsf{P}xw \wedge \mathsf{P}yw \rightarrow \mathsf{P}zw)) \qquad \textsf{Fusion}_a$$
$$(\text{P.11}_{\psi b}) \quad \psi xy \rightarrow \exists z(\mathsf{P}xz \wedge \mathsf{P}yz \wedge \forall w(\mathsf{P}wz \rightarrow \mathsf{O}wx \vee \mathsf{O}wy)) \qquad \textsf{Fusion}_b$$
$$(\text{P.11}_{\psi}) \quad \psi xy \rightarrow \exists z\forall w(\mathsf{O}wz \leftrightarrow \mathsf{O}wx \vee \mathsf{O}wy) \qquad \textsf{Fusion}$$

In a way, (P.11$_{\psi a}$) would seem the obvious choice, corresponding to the idea that the fusion of two objects is just their least upper bound relative to P. (See e.g. Bostock 1979, van Benthem 1983.) However, this condition is too weak to capture the intended notion of a mereological fusion. For example, with ψ construed as overlap, (P.11$_{\psi a}$) is satisfied by the model of Fig. 15.6, left: here the least upper bound of x and y is z, yet z hardly qualifies as something 'made up' of x and y since its parts also include a third, disjoint item w. In fact, it is a simple fact about partial orderings that among finite models (P.11$_{\psi a}$) is equivalent to (P.10$_\psi$), hence just as weak. By contrast, (P.11$_{\psi b}$) corresponds to a notion of fusion (to be found e.g. in Tarski 1929) that may seem too strong:

it rules out the model on the left of Fig. 15.6; but it also rules out the model on the right, which depicts a situation in which z may be viewed as an entity truly made up of x and y insofar as it is ultimately composed of atoms to be found either in x or in y. Of course, such a situation violates the strong supplementation principle (P.5), but that's precisely the sense in which (P.$11_{\psi b}$) is too strong: an anti-extensionalist might want to have a notion of fusion that does not presuppose strong supplementation. The formulation in (P.11_{ψ}) is the natural compromise: it is strong enough to rule out the model on the left, but weak enough to be compatible with the model on the right. This is, in fact, the formulation that best reflects the notion of fusion to be found in standard treatments of mereology, and in the sequel we shall mostly stick to it. Note, however, that *if* (P.5) holds, then (P.11_{ψ}) is equivalent to (P.$11_{\psi b}$). Moreover, it turns out that if the stronger complementation axiom (P.6) holds, then all of these principles are trivially satisfied in any domain in which there is a universal entity: in that case, regardless of ψ, the fusion of any two entities is just the complement of the difference between the complement of one minus the other. (Such is the strength of (P.6), a genuine cross between decomposition and composition principles.)

We can further strengthen these principles by considering infinitary bounds and fusions. For example, (P.10_{ψ}) can be generalized to a principle to the effect that any non-empty set of entities satisfying a suitable condition ξ has an upper bound. Strictly speaking there is a difficulty in expressing such a principle in a language without set variables. We can, however, achieve a sufficient degree of generality by relying on an axiom schema where classes are identified by open formulas. Since an ordinary first-order language has a denumerable supply of formulas, at most denumerably many sets (in any given domain) can be specified in this way. But for most purposes this limitation is negligible, as normally we are only interested in those sets of objects or regions that we *are* able to specify. Thus, the following axiom schema will do, where 'ϕ' is any formula in the language and 'ξ' expresses the condition in question:

(P.12_{ξ}) $\quad \exists w\phi w \wedge \forall w(\phi w \rightarrow \xi w) \rightarrow \exists z\forall w(\phi w \rightarrow \mathsf{P}wz)$

Likewise, the fusion axiom (P.11_{ψ}) can be strengthened as follows:

(P.13_{ξ}) $\quad \exists w\phi w \wedge \forall w(\phi w \rightarrow \xi w) \rightarrow \exists z\forall w(\mathsf{O}wz \leftrightarrow \exists v(\phi v \wedge \mathsf{O}wv))$

and similarly for (P.$11_{\psi a}$) and (P.$11_{\psi b}$). (The condition '$\exists w\phi w$' guarantees that 'ϕ' picks out a non-empty set, so there is no danger of asserting the unconditional existence of 'null entities'—a mereological fiction that we have already mentioned as implausible in the context of spatial ontology.) It can be checked that these generalized formulations include the corresponding finitary principles as special cases, taking 'ϕw' to be the formula '$(w = x \vee w = y)$' and 'ξw' the condition '$(w = x \rightarrow \psi wy) \wedge (w = y \rightarrow \psi xw)$'.

Finally, we get the strongest version of these composition principles by asserting them as axiom schemas that hold for *every* condition ξ, i.e., effectively, by foregoing any reference to ξ altogether. Formally this amounts in each case to dropping the second conjunct of the antecedent. For example, the following schema is the unrestricted version of (P.13$_\xi$), to the effect that every specifiable non-empty set of entities has a fusion:

(P.13) $\exists w\phi w \rightarrow \exists z\forall w(\mathsf{O}wz \leftrightarrow \exists v(\phi v \wedge \mathsf{O}wv))$ Unrestricted Fusion

The extension of EM obtained by adding every instance of this schema has a distinguished pedigree and is known as *General Extensional Mereology*, or GEM. It corresponds to the classic systems of Leśniewski and of Leonard and Goodman. In fact, it turns out that adding (P.13) to *MM* yields the same theory GEM, since (P.13) implies that every pair of overlapping things has a maximal common part (a product):

(38) $\mathsf{O}xy \rightarrow \exists z\forall w(\mathsf{P}wz \leftrightarrow (\mathsf{P}wx) \wedge \mathsf{P}wy))$

which, in turn, implies the equivalence between the weak supplementation principle (P.4) and the stronger (P.5) (Simons 1987, p. 31). This is by itself remarkable, for it might be thought that a composition principle such as (P.13) should be compatible with the rejection of a decomposition principle that is committed to extensionality. On the other hand, mereological extensionality is really a double-barreled thesis: it says that two wholes cannot be decomposed *into* the same proper parts but also, by the same token, that two wholes cannot be composed *out of* the same proper parts. So it is not entirely surprising that as long as PP is well behaved, as per (P.4), extensionality might pop up like this in the presence of substantive composition statements. (It is, however, noteworthy that it pops up as soon as (P.4) is combined with a seemingly innocent thesis such as (38), so the anti-extensionalist should keep that in mind.)

The intuitive idea behind all these principles is in fact best appreciated in the presence of extensionality, for in that case the relevant fusions must be unique. Just to confine ourselves to GEM, it is natural to consider the following fusion operator (where '$\imath$' is the definite descriptor[5]):

(39) $\Sigma x\phi x =_{df} \imath z\forall w(\mathsf{O}wz \leftrightarrow \exists v(\phi v \wedge \mathsf{O}wv))$ fusion

Then (P.13) and (P.5) can be simplified to a single axiom schema:

(P.14) $\exists x\phi x \rightarrow \exists z(z = \Sigma x\phi x)$ Unique Unrestricted Fusion

[5]We assume a classical logical background, with '$\imath$' defined as usual. Much of what follows, however, would also apply in case a free logic were used instead, with '$\imath$' assumed as part of the logical vocabulary proper. (See Simons 1991b for a free formulation of mereology.)

and the full strength of the theory can be seen by considering that its models are all closed under the following functors, modulo the absence of a null entity:

(40) $x + y =_{df} \Sigma z(\mathsf{P}zx \vee \mathsf{P}zy)$ sum
(41) $x \times y =_{df} \Sigma z(\mathsf{P}zx \wedge \mathsf{P}zy)$ product
(42) $x - y =_{df} \Sigma z(\mathsf{P}zx \wedge \mathsf{D}zy)$ difference
(43) $\sim x =_{df} \Sigma z\mathsf{D}zx$ complement
(44) $\mathcal{U} =_{df} \Sigma z\mathsf{P}zz$ universe

(Absent the null entity, $\mathcal{U}$ has no complement while products are defined only for overlapping pairs and differences for pairs that leave a remainder). Since these functors are the natural mereological analogue of the familiar Boolean operators, with fusion in place of set abstraction, it follows that the parthood relation axiomatized by GEM has the same properties as the set-inclusion relation. More precisely, it is isomorphic to the inclusion relation restricted to the set of all non-empty subsets of a given set, which is to say a complete Boolean algebra with the zero element removed—a fact that has been known since Tarski (1935).

There are other equivalent formulations of GEM that are noteworthy. For instance, it is a theorem of every extensional mereology that parthood amounts to inclusion of overlappers:

(45) $\mathsf{P}xy \leftrightarrow \forall z(\mathsf{O}zx \rightarrow \mathsf{O}zy)$

This means that in an extensional mereology 'O' could be used as a primitive and 'P' defined accordingly, and it can be checked that the theory defined by postulating (45) together with the unrestricted fusion principle (P.13) and the antisymmetry axiom (P.3) is equivalent to GEM. Another elegant axiomatization of GEM, due to an earlier work of Tarski (1929), is obtained by taking just the transitivity axiom (P.2) and the unique unrestricted fusion axiom (P.14).

1.4.2 Composition, existence, and identity. Arguably, the algebraic strength of GEM speaks in favor of this theory as an account of the structure of space, since it is rather intuitive (and common practice) to understand spatial regions in terms of sets of points mereologically related by set-inclusion. As a general theory of the mereology of all spatial entities, however, GEM reflects substantive postulates whose philosophical underpinnings are controversial. Indeed, all composition principles turn out to be controversial, just as the decomposition principles examined in Sec. 1.3. For, on the one hand, it appears that the weaker, conditional formulations, from (P.10$_\psi$) to (P.13$_\xi$), are just not doing enough work: not only do they depend on the specification of the limiting conditions expressed by the predicates 'ψ' and 'ξ'; they also treat such conditions as merely sufficient for the existence of bounds and fusions, whereas ideally we are interested in conditions that are both sufficient and

necessary. On the other hand, the stronger, unconditional formulations—most notably (P.13)—appear to go too far, not only because they tend to obliterate any difference between weak and strong supplementation, i.e., extensionality, but because they commit the theory to the existence of a large variety of *prima facie* implausible mereological composites. So what is the right way to go, at least insofar as we are interested in the compositional structure of the spatial realm?

Concerning the first sort of worry, one could of course strenghten every conditional formulation to a biconditional expressing both a necessary and sufficient condition for the existence of an upper bound, or a fusion. But then the question of how such conditions should be construed becomes crucial. For example, in connection with (P.10$_\psi$) we have mentioned the idea of construing 'ψ' as 'O', the rationale being that mereological overlap establishes an important connection between what may count as two distinct parts of a larger whole. However, as a necessary condition overlap is arguably too stringent. We may have misgivings about the existence of scattered entities consisting of spatially unrelated parts, such as the top of my body and the bottom of yours, or the collection of my umbrellas and your left shoes. But in some cases no such misgivings arise. In some cases it appears perfectly natural to countenance wholes that are composed of two or more disjoint entities—a bikini, the solar system, a printed inscription consisting of separate letter tokens (Cartwright 1975). More generally, intuition and common sense suggest that some and only some mereological composites exist, not all; yet it is doubtful whether the question of *which* composites exist—van Inwagen's 'special composition question'—can be answered successfully. Consider a series of almost identical mereological aggregates that begins with a case where composition appears to obtain (e.g., the body cells that currently make up my body) and ends in a case where composition would seem not to obtain (e.g., the same body cells after their relative distance has been gradually increased to a huge extent). Where should we draw the line? It may well be that whenever some entities compose a bigger one, it is just a brute fact that they do so (Markosian 1998b). But if we are unhappy with brute facts, if we are looking for a principled way of drawing the line so as to specify the circumstances under which the facts obtain, then the question is truly challenging. As Lewis (1986, p. 213) put it, no restriction on composition can be vague, since existence cannot be a matter of degree; but unless it is vague, it cannot fit the intuitive *desiderata*.

For these reasons, although the axiom of unrestricted fusion has been a major source of complaint since the early days of mereology (see again Lowe 1953 and Rescher 1955, with replies in Goodman 1956, 1958), it is a fact that most formally accomplished theories accept unrestricted composition principles of some sort. Apart from whatever algebraic considerations might motivate them, such principles suggest themselves as the only *non-arbitrary* ways of answering

the composition question. Besides, it might be observed that any complaints against such principles rest on psychological biases that have no bearing on how the world is actually structured. In the words of van Cleve (1986, p. 145), little would follow even if we did manage to come up 'with a formula that jibed with all ordinary judgments' about what counts as a unit and what does not, for such judgments need not have any ontological transparency.

All of this speaks in favor of (P. 13) and the like against their weaker, conditional formulations, providing also an answer to the second worry mentioned above: the *prima facie* ontological extravagance of a theory such as GEM is not by itself a sign that the theory has gone too far. There is, however, another worry that is worth mentioning in this connection, and this further worry concerns the ontological exuberance—if not the extravagance—of the theory. For even granting the impossibility of drawing a principled line between natural fusions and unnatural ones, one could still object that positing every conceivable fusion is utterly unjustified. Why should mereology be committed to the existence of all such things *over and above* their constituent proper parts?

There are two lines of response to this question. First, it could be observed that the ontological exuberance associated with the relevant composition principles is not substantive. This is obvious in the case of a modest principle in the spirit of (P.10$_\psi$), to the effect that entities of the right sort have an upper bound. After all, there are small things (my fingers) and large things (my body), and it is just a fact that the latter encompass the former. But the same could be said with respect to those stronger principles that require the large thing to be composed *exactly* of the small things—to be their mereological fusion. For one could argue that even a fusion is, in an important sense, nothing over and above its constituent parts. The fusion is just the parts 'taken together' (Lewis 1991, p. 81); it is the parts 'counted loosely' (Baxter 1988, p. 580); it is, effectively, the same portion of reality, which is strictly a multitude and loosely a single thing. This thesis, known in the literature as 'composition as identity', is by no means undisputed (see e.g. van Inwagen 1994, Yi 1999, Merricks 1999). Nonetheless it should be carefully evaluated in connection with any worry about the ontological exuberance of fusion principles. And if the thesis is accepted, then the charge of ontological extravagance loses its force, too. If a fusion is nothing over and above its constituent parts, and if the latter are all right, there can be nothing particularly extravagant in countenancing the former: it just is them.

Secondly, one may observe that the worry in question bites at the wrong level. If, given two entities, positing their sum were to count as further ontological commitment, then, given any mereologically composite entity, positing its proper parts should also count as further commitment. After all, every entity is distinct from its proper parts. But then the worry has nothing to do with the composition axioms; it is, rather, a question of whether there is any point in countenancing a whole along with its parts, or vice versa. And if the answer is

in the negative, then there seems to be no use for mereology *tout court*. From the point of view of the present worry, it would seem that the only truly parsimonious account would be one that rejects, not only *some*, but *all* logically admissible fusions—in fact, all mereological composites whatsoever. Philosophically such an account is defensible (see Rosen and Dorr 2002) and the corresponding axiom is compatible with M:

(P.15) $\mathsf{A}x$ Strong Atomicity

The following immediate corollary, however, says it all: nothing would be part of anything else and parthood would collapse to identity.

(46) $\mathsf{P}xy \leftrightarrow x = y$

(This account is known as mereological *nihilism*, in contrast to the mereological *universalism* expressed by (P.13); see van Inwagen 1990, pp. 72ff.)

In recent years, further worries have been raised concerning mereological theories with non-trivial composition principles—especially concerning the full strength of GEM. It has been argued that unrestricted composition does not sit well with certain intuitions about persistence through time (van Inwagen 1990, 75ff), that it requires every entity to necessarily have the parts it has (Merricks 1999), or that it leads to paradoxes similar to the ones afflicting naive set theory (Bigelow 1996). Such arguments are still the subject of on-going controversy and a detailed examination is beyond the scope of this chapter. Some discussion of the first point, however, is already available in the literature: see especially Rea (1998), McGrath (1998), and Hudson (2001, pp. 93ff). Hudson (2001, pp. 95ff) also contains a discussion of the last point.

1.5 The problem of vagueness

Let us conclude this discussion of mereology by considering a question that is not directly related to specific mereological principles but, rather, to the underlying notion of parthood that mereology seeks to systematize. All the theories examined so far, from M to GEM, assume that parthood is a perfectly determinate relation: given any two entities x and y, there is always a definite fact of the matter as to whether or not x is part of y. However, it may be argued that this is a simplification. Perhaps there is no room for vagueness in the idealized mereology of pure space, but what about the real world? Think of a cloud, a forest, a pile of trash. What parts do they have, exactly? What are the mereological boundaries of a desert, a river, a mountain? Some stuff is positively part of Mount Everest and some stuff is positively not part, yet there is borderline stuff whose mereological relationship to Everest seems indeterminate. Even living organisms may, on closer look, give rise to vagueness issues. Surely John's body comprises his heart and does not comprise mine. But what about

Figure 15.7. Objects with indeterminate parts (in grey).

the candy he is presently chewing: Is it part of John? Will it be part of John only after he swallowed it? After he started digesting it? After he digested it completely?

In the face of such examples, it might be thought that the conceptual apparatus on which M and its extensions are based is too rigid. It might be thought that the world includes various sorts of vague entities, and that relative to such entities the parthood relation need not be fully determined (van Inwagen 1990: Ch. 13, Parsons and Woodruff 1995). There are, in fact, various ways one could seek greater flexibility. One could leave everything as is but change the underlying logic (and semantics), for instance by allowing statements of the form '$\mathsf{P}xy$' to receive no determinate truth-value (as in Tye 1990), or to receive truth-values that are intermediate between classical truth and falsity (as in Copeland 1995). Or one could change the very basic apparatus of mereology, replacing the 'part of' predicate with a new primitive 'part of *to a degree*': this is, for example, the approach that led to the development of Polkowsky and Skowron's (1994) 'rough mereology', where parthood undergoes a fuzzification parallel to the fuzzification of membership in Zadeh's (1965) fuzzy set theory. No matter how exactly one proceeds, obviously many among the principles discussed above would have to be reconsidered, not because of what they say but because of their classical, bivalent presuppositions. For example, the extensionality theorem of EM, (31), says that composite things with the same proper parts are identical, and this would call for qualifications: the model in Fig. 15.7, left, depicts x and y as non-identical by virtue of their having distinct determinate parts; yet one might prefer to describe a situation of this sort as one in which the identity between x and y is itself indeterminate, since it is indeterminate whether they really have distinct parts. Conversely, the model on the right depicts x and y as non-identical in spite of their having the same determinate proper parts; yet again one might prefer to suspend judgment owing to the indeterminacy the middle element.

That there are vague objects in this sense, however, i.e., objects whose mereological composition may to some extent be objectively indeterminate, is all but obvious. Surely a statement such as

(47) x is part of Everest

may lack a definite truth-value, if x lies somewhere in the borderline area. But—it could be argued—this need not be due to the way the world is. The

indeterminacy of (47) may be due exclusively to semantic factors—not to the vagueness of Everest but to the vagueness of 'Everest'. When the members of the Geodetic Office of India baptized the mountain after the name of their British founder, they simply did not specify exactly which parcel of land they were referring to (or which parcel of land constituted the mountain they meant to name).[6] The referent of their term was vaguely fixed and, as a consequence, the truth conditions of a statement such as (47) are not fully determined; yet this is not to say that the stuff out there is mereologically vague. Each one of a large variety of slightly distinct parcels of land has an equal claim to the vaguely introduced name 'Everest', and each such thing has a perfectly precise mereological structure. To put it differently, a statement such as

(48) It is indeterminate whether x is part of Everest

admits of a *de re* reading, as in (48a), but also of a *de dicto* reading, as in (48b):

(48a) Everest is a y such that: it is indeterminate whether x is part of y.
(48b) It is indeterminate whether: Everest is a y such that x is part of y.

The first reading corresponds to the initial thought, to the effect that Everest's parts are indeed indeterminate, with the consequence that mereology ought to be revised as seen above. The second reading, by contrast, corresponds to the idea that it is the semantics of 'Everest' that is indeterminate, and there is no reason to suppose that this is due to some objective deficiency in the parthood relation—hence no reason to require revisions in the apparatus of mereology itself. (Ditto for the other cases mentioned above. The reason why it's indeterminate whether a certain molecule is part of a cloud, a tree part of a forest, or the candy part of John, is not that such things are mereologically indeterminate; rather, on a *de dicto* understanding the indeterminacy lies entirely in our words, in the terms we use to pick out such things from a multitude of slightly distinct but perfectly determinate potential referents.)

If the semantic conception is accepted, then, the problem of vagueness dissolves. Or rather: it ceases to be an issue for mereology and it becomes a problem for semantics broadly understood—a problem that manifests itself in many contexts besides those under consideration. (How much money do you need to be *rich*? How slowly can you *run*? How *late* can I call you?) Again, there are many things one could do at this point. A favored option is afforded

[6]That mountains are just parcels of land is, of course, a substantive assumption: an anti-extensionalist may want to deny it, as with Tibbles and the relevant amount of feline tissue (Sec. 1.3.2). On the ontology of topographic entities, see e.g. Smith and Mark (2003).

by so-called supervaluational semantics, whose first application to vagueness can be traced back to Fine (1975). According to such semantics, the truth-value of a sentence involving vague terms is a function of its truth-values under the admissible precisifications of those terms: the sentence is true if it is true under every precisification, false if false under every precisification, and indeterminate otherwise. Thus, if x is in the borderline area, then the indeterminacy of (47) is explained by the fact that among the many admissible ways of precisifying the term 'Everest', some would pick out a referent that extends far enough to include x among its parts whereas others would not, which is to say that (47) would be true on some but not all precisifications. By contrast, if x were clearly part of Everest given the way the name is used in ordinary circumstances, or if it were clearly not part of Everest, then (47) would have a definite truth-value, for every precisification would yield the same response (always true and always false, respectively). We need not go into the details here. But three things are worth noting.

First, none of this will have any impact on the mereological axioms considered so far. For those axioms are expressed as (implicitly) universally quantified formulas involving no singular terms except for variables, and variables are not the sort of expression that can suffer from the phenomenon of vagueness. Variables range over all entities included in the domain of quantification and pick out their values independently of any vagueness that may affect the non-logical vocabulary.

Second, any model that satisfies a given axiom or theorem satisfies also any substitution instance that can be obtained by replacing one or more variables with corresponding names or descriptive terms. For example, the following sentence is a substitution instance of (P.1):

(49) Everest is part of Everest.

and it is easily verified that a supervaluational semantics will make (49) true in every model of M. For insofar as reflexivity is meant to hold for every entity in the domain, the truth of (49) is guaranteed no matter which entity we elect as the referent of 'Everest'. Likewise, the following is a substitution instance of (31), the extensionality theorem of EM:

(50) As long as they are non-atomic, Everest and Sagarmatha are the same if and only if they have the same proper parts.

('Sagarmatha' is the Nepalese translation of 'Everest', though there is no guarantee that both names admit of the same precisifications.) Again, it is easily verified that a supervaluational semantics will make (50) true in every EM model. For no matter how we precisify the terms 'Everest' and 'Sagarmatha' by tracing a precise boundary around their referents, the extensionality of part-

hood will guarantee that the referents coincide just in case their proper parts coincide too.

Finally, it is worth emphasizing that a supervaluational semantics is perfectly adequate to classical logic (Fine 1975, McGee 1997, Varzi 2001). For example, although it does not obey to the semantic principle of *bivalence*, as with various instances of

(51) Either 'x is part of Everest' is true or 'x is part of Everest' is false,

it certainly satisfies the logical law of *excluded middle*: any instance of

(52) Either x is part of Everest or x is not part of Everest

is bound to be true, for it is true on any precisification of 'Everest'. Things would change, however, if the language were enriched by adding an explicit sentential operator to express indeterminacy. In that case, the following principle would give expression to the assumption that parthood admits of no objective borderline cases:

(P.16) It is determinate whether $\mathsf{P}xy$, Determinacy

though it is obvious that this principle may have invalid instances as soon as 'x' or 'y' is replaced by a vague singular term such as 'Everest', as in (48). At the moment, the logic of determinacy operators is an open area of research, especially in the context of a supervaluational semantics. (See Keefe 2000, §7.4 and references therein.) It is, however, an important tool that any good semantic theory of vagueness should countenance. And it is bound to play a significant role in any application of the theory to mereology and spatial reasoning broadly understood.

2. Philosophical issues in topology

Let us now move on to the second major ingredient of a comprehensive theory of spatial reasoning—topology. There are many reasons for this move, but the main one is simply this: one need go beyond the bounds of a pure theory of parthood to come out with a true theory of parts and *wholes*. For as we have already mentioned, mereology by itself cannot do justice to the notion of a whole (a one-piece, self-connected whole such as a stone or a whistle, as opposed to a scattered entity made up of several disconnected parts, such as a bikini or a broken glass). Parthood is a relational concept, wholeness a global property, and the latter just runs afoul of the former.

Whitehead's early attempts to characterize his ontology of events, as presented at length in his *Enquiry* (1919) and in *The Concept of Nature* (1920), exemplify this difficulty most clearly. The mereological system underlying

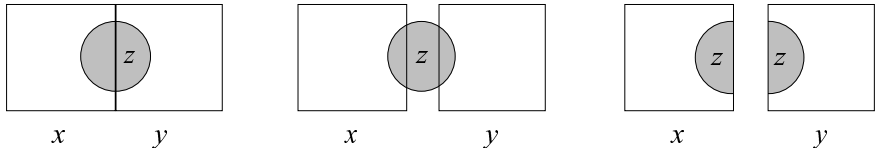


Figure 15.8. A connected sum (left) and two disconnected sums (middle, right).

Whitehead's ontology was not meant to admit of arbitrary wholes, but only of wholes made up of parts that are 'joined' or connected to one another—specifically, finitary sums of such parts. Thus, Whitehead was working with a composition principle patterned after (P.11$_\psi$), in fact with the corresponding biconditional, with 'ψ' understood as a predicate expressing the relevant relation of connection. And Whitehead's characterization of this relation was purely mereological:[7]

(53) $\psi xy =_{df} \exists z(\mathsf{O}zx \wedge \mathsf{O}zy \wedge \forall w(\mathsf{P}wz \rightarrow \mathsf{O}wx \vee \mathsf{O}wy))$

Looking at the spatial patterns in Fig. 15.8, we can see how this definition is intended to work. What distinguishes the connected sum $x + y$ on the left from the disconnected sum in the middle? Well, in the former case it is easy to find regions, such as z, that overlap both x and y without outgrowing the sum—regions that lie entirely *within* $x + y$. By contrast, in the middle pattern it would seem that every z overlapping both x and y will also overlap their complement—the entity that *surrounds* $x + y$. Thus, only the left pattern satisfies the condition expressed by 'ψ'; the pattern in the middle violates it. However, Fig. 15.8 also shows why this account is defective. For nothing guarantees that the item z overlapping two 'joined' items x and y be itself in one piece, so the pattern on the right depicts two entities that satisfy the condition expressed by 'ψ', too. Yet this is a case where we should like to say that x and y are *not* connected. Of course, Whitehead would disqualify the counterexample because his ontology does not contain any disconnected zs—but this is plain circularity. The account works on the assumption that only self-connected entities can inhabit the domain of discourse, yet that is precisely the assumption that (53) is meant to characterize.

These considerations apply *mutatis mutandis* to other attempts to subsume topological connectedness within a bare mereological framework (see e.g. Bostock 1979, Needham 1981, Ridder 2002). Nor is this exclusively an ontological concern. These limits show up in any attempt to account for a number of important spatial concepts besides connectedness, such as the distinction

[7]The definition below corresponds to the formulation given in Whitehead (1920, p. 76). Whitehead's earlier definition (1919, p. 102) is slightly different but essentially equivalent.

between a completely interior part and a tangential part that is connected with the exterior, of the difference between an open entity and a closed one. All of these—and many others indeed—are relations that any theory concerned with the spatial structure of the world should supply and which cannot, however, be defined directly in terms of plain mereological primitives.

2.1 'Contact' and connection

It is here that topology comes into the picture. The connection relation that Whitehead was seeking to characterize is a topological relation. And if it cannot be defined in mereological terms, it must be formally treated on independent grounds.

Before looking at how this can be done, it is important again to begin with a couple of terminological caveats. As with 'part', the term 'connection', and cognate terms such as 'contact' and 'touching', have different meanings in ordinary language, only some of which correspond to the intended relation. Most notably, in ordinary language we do not draw a clear distinction between a truly topological notion of connection and a merely metric notion of contact. Consider:

(54) The handle is attached to the mug.
(55) The table is touching the wall.

The relation exemplified by (54) is topological: the handle and the rest of the mug form a unitary whole. For practical purposes there may be room for free rein, depending on whether the handle is glued to the rest of the mug (what Galton 2000, §4.2, calls 'adhesion') or truly continuous with it ('cohesion'), but either way there is an obvious sense in which we are dealing with a single, one-piece object. By contrast, the relation exemplified by (55) is not topological but metric: the table is so close to the wall that we are inclined to say they are connected. If space were discrete, this might be the right thing to say. But if space is dense, as we may plausibly assume, then the surfaces of two bodies can never truly be connected, short of overlapping: there will always be a narrow gap separating them. The narrower this gap, the easier it is to disregard it for practical purposes, but genuine topological connection can only obtain when the gap is reduced to zero. We shall see in Sec. 2.5.1 that some interesting topological relations may be introduced to capture at least some uses of the metric relation of contact. Overall, however, metric relations cannot be squeezed into the conceptual apparatus of topology without the help of strongly simplifying assumptions on the structure of space. (In this sense, a diagram such as Fig. 15.8, left, is ambiguous, since one might think of the two squares x and y as being merely close to each other, as when we draw a picture of a table against a wall.)

A related issue concerns the distinction between connection patterns that involve a single point of contact, as in (56), or an extended boundary portion, as in (57), if not of an entire boundary, as in (58):

(56) Colorado is connected to Arizona.
(57) France is connected to Germany.
(58) The Vatican is connected to Italy.

To some extent this is a matter of convention. We may disregard (56) as irrelevant, or we may treat it as an acceptable case.[8] Since most theories go for the second option, we shall follow them on this score. As it turns out, under suitable conditions one can on such basis draw all the relevant distinctions (Sec. 2.5), so the choice proves to be a convenient one.

Finally, let us just mention the fact that topological connection is, in a way, an idealized relation. Physically, as we know, the world consists of objects that are not continuous (or dense) in the relevant sense, and speaking of their boundaries is like speaking of the 'flat top' of a fakir's bed of nails (Simons 1991a, p. 91). Physically, a mug is just a swarm of subatomic particles whose exact shape and extension involves the same degree of arbitrariness as a mathematical graph smoothed out of scattered data. In this sense, the intuition behind the claim that (54) provides a good example of topological connection betrays a naive conception of the mid-size world. However, this is not to say that topology is inadequate as a tool for practical spatial reasoning. For, on the one hand, we are generally interested in describing the spatial structure of the world precisely insofar as objects are conceptualized as finite chunks of dense matter with closed, continuous boundaries. Even if talk of boundaries and contact were deemed unsuited to the ontology of the physical sciences, one would therefore need it when it comes to the dense entities carved out by ordinary discourse and to the spatial regions that these occupy. On the other hand, the geographic examples in (56)–(58) illustrate that at least *some* entities countenanced by common sense measure up to the strict standards of topological connection. We do want to say that a geopolitical unit occupies a region of space that is strictly dense in the relevant sense, even though the underlying territory may consist of material stuff that on closer inspection is best described as a gerrymandered aggregate of zillions of disconnected subatomic particles.

2.2 Basic principles and definitions

We are now in a position to take a closer look at the idea of a topological extension of mereology—an extension that would take us beyond the prospects of

[8]This may also depend on context. In chess, for example, the choice depends on whether we are a rook or a bishop.

a pure theory of parthood to deliver a genuine theory of parts and *wholes*. To this end, let us expand our formal language by adding a second distinguished predicate constant, 'C', to be understood intuitively as the relation of topological connection. The question of how mereology can actually be expanded to a richer *part–whole* theory may then be addressed by investigating how a P-based mereological system of the sort outlined in Sec. 1 can be made to interact with a C-based topological system.

Again, we may distinguish for this purpose 'lexical' from substantive postulates for 'C', regarding the former as embodying a set of minimal prerequisites that any system purporting to explicate the meaning of the concept of 'connection' must satisfy. And a natural starting point is to assume that such lexical principles include at least the twofold requirement that 'C' be reflexive and symmetric:

(C.1) $\mathsf{C}xx$ Reflexivity
(C.2) $\mathsf{C}xy \rightarrow \mathsf{C}yx$ Symmetry

There is little room for controversy concerning the intuitive adequacy of (C.1)–(C.2), provided that we take 'C' to express, not just the relation of *external* connection that may obtain between two disjoint entities that share a common boundary (in some intuitive sense to be made precise), but the relation of connection that may obtain between any two entities that share *at least* a boundary. In this sense, mereological overlap qualifies as connection, too. We shall come back to this idea shortly. First, let us note that 'C' need not, on the intended interpretation, express a transitive relation. France is connected to Germany and Germany to Poland, but France and Poland are not connected—they do not share any common boundary. We can, however, consider a notion of connection that captures the fact that France is connected to Poland *by* Germany. De Laguna (1922), a forerunner in the area of qualitative topological reasoning, actually based his account on a three-place primitive corresponding to this relation.[9] In terms of 'C', De Laguna's primitive is easily defined:

(59) $\mathsf{BC}xyz =_{df} \mathsf{C}xz \wedge \mathsf{C}zy$ By-Connection

and we can accordingly introduce the desired notion of (possibly) indirect or mediate connection as follows:

(60) $\mathsf{MC}xy =_{df} \exists z \mathsf{BC}xyz$ Mediate Connection

By an obvious generalization, we can also define:

[9]Strictly speaking, De Laguna's primitive is interpreted as 'x can be connected to y by z', so it involves a modal ingredient. For a formal treatment, see Giritli (2003).

(61) $\mathsf{MC}^n xy =_{df} \exists z_1 ... z_n (\mathsf{C}xz_1 \wedge ... \wedge \mathsf{C}z_n y)$ n-Connection

It follows immediately from (C.1) and (C.2) that each MC^n is reflexive and symmetric, and the union of all such relations is transitive. In the absence of further principles, however, e.g., principles guaranteeing the existence of an entity connected to all the intermediate links, such a transitive union cannot be defined in the object language—unless we allow for quantification over positive integers:

(62) $\mathsf{TC}xy =_{df} \exists n \mathsf{MC}^n xy$ Transitive Connection

Now, let T be the first-order theory defined by the two basic axioms (C.1) and (C.2), in analogy with the theory M defined by the basic mereological axioms (P.1)–(P.3). T is, of course, extremely weak and a lot will have to be added before we can say we have an interesting topology. In particular, a model of T can be obtained simply by interpreting 'C' as mereological overlap, and what further principles should be added to T so as to distinguish C from O is precisely one of the questions a good topological theory is meant to answer. For instance, should one assume that connection is extensional, i.e., that things that are connected exactly to the same entities are identical? Should one assume that any two connected entities satisfy at least some form of Whitehead's account in (53)? Or consider the binary relation defined by

(63) $\mathsf{E}xy =_{df} \forall z(\mathsf{C}zx \rightarrow \mathsf{C}zy)$ Enclosure

It follows from (C.1)–(C.2) that this relation is reflexive and transitive, and if C is extensional, than E is also antisymmetric—a partial ordering. Should one assume this relation to satisfy any analogues of the axioms for parthood? For each mereological predicate defined in Sec. 1 using 'P' one could now introduce a corresponding topological predicate using 'E' instead. Should one assume any corresponding axioms?

As it turns out, it is difficult to answer these questions in an abstract setting (see Cohn and Varzi 2003). Obviously, much depends on how exactly 'C' is interpreted, and that in turn may depend on how one thinks 'C' and 'P' should interact. Rather than pursuing these questions in isolation, then, let us proceed immediately to examining the main options for combining mereology and topology.

2.3 Bridging principles

The simplest option is just to append the T-axioms to our preferred mereological theory, X, to obtain a corresponding 'mereotopology' $X + T$, which can then be strengthened by supplying further axioms for 'C'. However, this would be of little interest unless one also adds some mixed principles to establish an explicit 'bridge' between X and T.

2.3.1 Parts and wholes. There is one sort of bridging principle that most theories, if not all, accept: it centers around the intuition that no matter how P and C are fully characterized, they must be related in such a way that a whole and its parts are tightly connected. Here are three ways one can try to capture this intuition:

(C.3$_a$)	$\mathsf{P}xy \rightarrow \mathsf{C}xy$	Integrity
(C.3$_b$)	$\mathsf{O}xy \rightarrow \mathsf{C}xy$	Unity
(C.3)	$\mathsf{P}xy \rightarrow \mathsf{E}xy$	Monotonicity

The first principle, (C.3$_a$), is perhaps the most immediate: just as everything is connected to itself by (C.1), everything must be connected to its constitutive proper parts. This is not to say that the parts must all be connected to one another: the two main parts of a bikini are not. But they are, in an obvious sense, connected to their sum; they are detached from each other but not from the whole bikini.

As it stands, however, (C.3$_a$) is extremely weak. It doesn't even capture the idea that if something is part of *two* things, then those things are thereby connected. This is not to say that they are connected *by* that common part, in the sense defined in (59); they are connected *because* of that common part. In other words, if sharing a common boundary is to count as sufficient for connection, then *a fortiori* sharing a common part ought to be sufficient, too. It is in this sense that overlap is to be regarded a special (and somewhat trivial) case of connection. And the second principle, (C.3$_b$), makes this explicit.

(C.3$_b$) is stronger than (C.3$_a$), since parthood implies overlap. Moreover, since the converse need not hold (on pain of trivializing the notion of connection), (C.3$_b$) provides the intuitive grounds for defining a non-trivial notion of external connection, or touching, that can only hold between mereologically disjoint entities:

(64) $\mathsf{EC}xy =_{df} \mathsf{C}xy \wedge \mathsf{D}xy$ External Connection

(Note that this relation is still symmetric, but not reflexive; it is actually irreflexive, due to the irreflexivity of D.) This is an important notion, which makes all the difference between mereology and mereotopology. Yet (C.3$_b$) is still too weak to capture the fundamental intuition that we are after. For while this principle guarantees that overlapping a part is sufficient for being connected to the whole, it doesn't secure that *touching* a part is also sufficient. Surely something can touch a mug (say) just by touching its handle. So it is only with the third principle, (C.3), that we get a plausible formulation of the basic idea. Connection, if it is to behave properly, must be monotonic with respect to parthood.

It is easily checked that (C.3) implies (C.3$_b$), hence (C.3$_a$), so let us just focus on (C.3), and let us call MT (for *Minimal (mereo)Topology*) the corresponding extension of T.[10] Whether (C.3) is to be classified as a 'lexical' principle may be controversial and will depend, in an obvious sense, on the underlying axioms for 'P'. Nonetheless, the principle itself is part of virtually every mereotopological theory in the literature, either as an axiom (Varzi 1996a, Donnelly 2004) or as a theorem. And although MT is still far from providing an adequate characterization of the relation of topological connection, it provides the basis for the definition of a number of important spatial relations which, like EC, cannot be distinguished within a purely mereological setting. In particular, we can now express the difference between a proper part that lies entirely within the interior of the whole and a proper part that is connected with the exterior:

(65) $\mathsf{IPP}xy =_{df} \mathsf{PP}xy \wedge \forall z(\mathsf{C}zx \rightarrow \mathsf{O}zy)$ Interior PP

(66) $\mathsf{TPP}xy =_{df} \mathsf{PP}xy \wedge \neg\mathsf{IPP}xy$ Tangential PP

(67) $\mathsf{EPE}xy =_{df} \mathsf{PE}xy \wedge \forall z(\mathsf{C}zy \rightarrow \mathsf{O}zx)$ Exterior PE

(68) $\mathsf{TPE}xy =_{df} \mathsf{PE}xy \wedge \neg\mathsf{IPE}xy$ Tangential PE

Note that, given (64), interior parts satisfy the following:

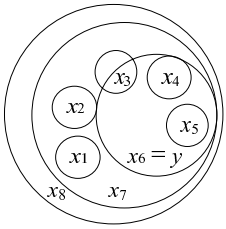
(69) $\mathsf{IPP}xy \leftrightarrow (\mathsf{PP}xy \wedge \neg\exists z(\mathsf{EC}zx \wedge \mathsf{EC}zy))$

Thus, tangential parts are those parts that reach far enough to touch something with which the whole itself is just in touch. Similarly for proper extensions. The diagram in Fig. 15.9 indicates how these predicates may represent a genuine addition to the mereological vocabulary introduced in (19)–(23) and illustrated in Fig. 15.1. We shall see in Sec. 2.4 that this diagram may actually be misleading, owing to the delicate role played by boundaries in the proper understanding of the connection relation; but for the moment we take the intuitive, geometric interpretation of the diagram to be adequate enough to serve its purpose.

2.3.2 Parthood vs. enclosure. Things begin to be controversial as soon as we consider the possibility of stronger bridging principles. Consider again the three principles above. Clearly the converse of Integrity, (C.3$_a$), is unacceptable. And unacceptable is also the converse of Unity, (C.3$_b$), for then connection would collapse on the relation of mereological overlap and the definitions in (64)–(68) would lose their intuitive appeal. On the other hand, the converse of the Monotonicity principle (C.3) is not obviously unreasonable:

(C.4) $\mathsf{E}xy \rightarrow \mathsf{P}xy$ Converse Monotonicity

[10] In fact, the result of adding (C.3) to T yields a slightly redundant theory; a more elegant formulation can be obtained by dropping (C.2) and replacing (C.3) with the following variant: (C.3′) $\mathsf{P}xy \wedge \mathsf{C}xz \rightarrow \mathsf{C}zy$.



	U$x_n y$	EC$x_n y$	O$x_n y$	TPP$x_n y$	IPP$x_n y$	EQ$x_n y$	TPE$x_n y$	EPE$x_n y$
$n = 1$	+	–	–	–	–	–	–	–
$n = 2$	+	+	–	–	–	–	–	–
$n = 3$	+	–	+	–	–	–	–	–
$n = 4$	+	–	+	+	–	–	–	–
$n = 5$	+	–	+	–	+	–	–	–
$n = 6$	+	–	+	–	–	+	–	–
$n = 7$	+	–	+	–	–	–	+	–
$n = 8$	+	–	+	–	–	–	–	+

Figure 15.9. Basic mereotopological relations. (Shaded cells indicate connection; darker shading stands for parthood.)

This says that a sufficient condition for one thing to be part of another is that whatever is connected to the former is also connected to latter. This sounds intuitive, and several authors would actually include (C.4) as a further bridging principle on top of MT. Indeed, a principle along these lines may already be found in Whitehead's latest work, *Process and Reality* (1929)—a remarkable fact, since the conjunction of (C.3) and (C.4) yields a biconditional that would allow one to define parthood in terms of connection:

(70) $\mathsf{P}xy \leftrightarrow \mathsf{E}xy$

Thus, if (53) turned out to be a defective attempt to reduce topological concepts to purely mereological ones, (C.4) (together with (C.3)) reflects a reductionist attempt in the opposite direction, to the effect that mereological concepts can be defined in terms of purely topological ones. And although few have followed Whitehead on the first route, it is a fact that many authors have taken the second strategy into serious consideration. Clarke (1981) provides the most influential example, and the so-called 'Region Connection Calculus' originated with Randell *et al.* (1992) is the best case in point when it comes to theories designed specifically for applications to spatial reasoning (see also Gotts *et al.* 1996 and Cohn *et al.* 1997). So the question deserves close scrutiny: Is (C.4) a reasonable addition to the basic postulates of MT?

Never mind the fact that working with just one primitive may be mathematically attractive. As it turns out, it is equally possible to rely on a single primitive even in the absence of (C.4). For instance, one can rely on the ternary relation 'x and y are connected parts of z' (Varzi 1994). Writing this as '$\mathsf{CP}xyz$', one could define 'P' and 'C' as follows:

(71) $\mathsf{P}xy =_{df} \exists z\ \mathsf{CP}xzy$
(72) $\mathsf{C}xy =_{df} \exists z\ \mathsf{CP}xyz$

and then go on to develop a mereotopological theory based on the relative irreducibility of these two predicates. (Note that (71) and (72) only carry

Figure 15.10. Two counterexamples to converse monotonicity.

minimal presuppositions: that parthood be reflexive and that connected entities have an upper bound.) So the issue is not formal economy—the use of a single primitive. It is, rather, conceptual economy. In the extension of MT obtained by adding (C.4)—henceforth RMT, for *Reductive Mereotopology*—the *notion* of parthood is fully subsumed under that of connection, and the limits of mereology are overcome by turning the original problem upside down: parthood cannot deliver the full story, but connection can. However, there are at least two worries here.

The first worry concerns the material adequacy of the reduction. As Masolo and Vieu (1999) have observed (but the point goes back to Randell *et al.* 1992: §5.1), (C.4) appears to have implausible consequences if the domain contains entities with atomic proper parts. Consider an extended region (or object) x, and let y be x minus an atomic part z (Fig. 15.10, left). On any reasonable understanding of 'C', everything connected to x is connected to y, since z is connected to both. So (C.4) would force x to be part of y. Yet, intuitively, x should count as an extension of y: it is bigger, it contains z, it contains y as a further *proper* part. Things get worse if we consider that (C.4) forces z itself to count as part of y, since z is connected to y and anything else is connected to z only if it overlaps y. Yet y was defined as x *minus* z. Of course, such models would be ruled out if RMT were strengthened by adding an atomlessness postulate such as (P.7). But this is precisely the point: (C.4) does not merely reinforce the bridge between P and C; it actually embodies more substantive views about the mereological structure of space. Besides, even the atomless variant $\bar{A}RMT$ would be open to counterexamples. For we have the same sort of problem if we suppose that z is a non-atomic proper part of x with no interior proper parts of its own (Fig. 15.10, right). To rule out this model, a stronger assumption than (P.7) would be needed, corresponding to the thesis that everything has *interior* proper parts:

(C.5) $\exists x \mathsf{IPP} xy$ Boundarylessness

In fact, it is precisely with the help of an axiom like (C.5) that we can give expression to a Whiteheadian, boundary-free conception of space (see below, Sec. 2.4.3). However, this is a controversial conception. Why should our

analysis of parthood force upon us a rebuttal of the boundary concept? Why should we assume (C.5) in order to ensure a coherent implementation of a basic bridging principle like (C.4)?

The second worry is more general. Consider an object and the stuff it is made of—for instance, a statue and the corresponding amount of clay. As we have seen, few would regard the relationship of material constitution that holds between such entities as a case of proper parthood. And there are many philosophers for whom constitution is not identity (improper parthood) either: see again the papers in Rea 1997. This is by no means a pacific thesis, but never mind. The point is simply that the relationship between the clay and the statue is not obviously an instance of parthood. Yet, on any plausible understanding of 'C', whatever is connected to the clay is bound to be connected to the statue, too, so (C.4) would immediately settle the issue: the clay *is* part of the statue. Indeed, since it is equally plausible to suppose that the same applies in the opposite direction—whatever is connected to the statue is connected to the clay—(C.4) implies that the statue and the clay are one and the same thing. And this is a substantive tenet which, as such, ought not to be built into the basic apparatus of mereotopology *at the outset*.

This worry is perhaps best appreciated by noting that the following theorem is an immediate consequence of (C.1)–(C.4), provided P satisfies the basic mereological axioms (P.1)–(P.3):

(73) $\forall z(\mathsf{C}zx \leftrightarrow \mathsf{C}zy) \rightarrow x = y$

In fact, (P.1) and (P.2) (reflexivity and transitivity) are derivable from (70), and RMT turns out to be equivalent to the theory defined by taking (73) as an axiom along with (C.1) and (C.2). (That is actually the customary axiomatization since Clarke 1981; see also Biacino and Gerla 1991.) Now, with parthood construed as enclosure, (73) is nothing but the antisymmetry principle (P.3). Yet (73) does not merely assert the antisymmetry of parthood; it says that connection is extensional—that different things cannot connect to the same things. And this is just as controversial as the thesis that parthood is extensional.

It could be replied that the analogy with mereological extensionality is in fact helpful, since the original arguments in support of (P.5) (Sec. 1.3.2) could now be offered on behalf of (C.4). Indeed, if a statue and the clay are construed as things that, at some level of decomposition, share the same constituents—e.g., the same molecules—the question of whether constitution is identity just *is* the question of whether parthood is extensional. However, the worry does not only apply to cases of material constitution. Consider a shadow cast onto a wall. The shadow is not part of the wall, yet anything connected to the shadow is—arguably—connected to the wall. Or consider a stone inside a hole. The stone is not part of the hole, yet one could argue that anything connected to the stone is connected to the hole. Broadly speaking, the problem arises as soon as

we allow for the possibility that distinct entities occupy the same space (Casati and Varzi 1999). That this is a real possibility is by itself contentious and is one of the questions to be addressed by an explicit theory of location (Sec. 3). But precisely for this reason, ruling it out on mereotopological grounds would be utterly inappropriate.

This last point is particularly worth stressing, for it shows once again how the choice of a suitable set of principles may depend crucially on whether we are interested in a theory aimed at modeling the domain of all spatial entities or just a domain of pure spatial regions. The worry mentioned above arises forcefully in the context of theories of the first sort. It does, however, lose its force in relation to the second sort of theory, since two regions cannot overlap spatially without overlapping mereologically. Now, it is a fact that most authors committed to (C.4) have been working on such a narrower understanding of their theory. Whitehead's own account was explicit in this regard: in the theory put forward in *Process and Reality* (as in Clarke's 1981 reformulation), the field of C was meant to consist exclusively of spatial regions. On the other hand, it is also a fact that a major motivation for developing a theory of this sort has been the assumption that connection thus understood is all that matters for practical purposes. For one can always treat the relation of connection as the 'shadow' (in De Laguna's 1922, p. 450 apt terminology) of the relation of physical contact or overlap that may obtain between actual, concrete entities. In other words, such theories have typically been developed on the assumption that the following principle provides the necessary and sufficient link between the mereotopology of pure space and the mereotopology of spatial entities broadly understood:

(74) x is connected to y if and only if the region occupied by x is connected to the region occupied by y.

If so, however, then the problems mentioned above resurface even for theories of this sort. For (74) will deliver an acceptable account if, and only if, spatial co-location is regarded as metaphysically impossible.

2.3.3 Self-connectedness. There are other ways of supplementing MT with bridging principles that go beyond (C.3). In particular, consider again Whitehead's early attempts to characterize topological connection in terms of parthood, i.e. (53), with 'ψ' understood as 'C'. As a *definition*, this was found defective. However, one may certainly consider adding the corresponding biconditional as an axiom—or at least adding one of the two conditionals:

(C.6) $\mathsf{C}xy \rightarrow \exists z(\mathsf{O}zx \wedge \mathsf{O}zy \wedge \forall w(\mathsf{P}wz \rightarrow \mathsf{O}wx \vee \mathsf{O}wy))$ Left Join

(C.7) $\exists z(\mathsf{O}zx \wedge \mathsf{O}zy \wedge \forall w(\mathsf{P}wz \rightarrow \mathsf{O}wx \vee \mathsf{O}wy)) \rightarrow \mathsf{C}xy$ Right Join

Would this be a good way of tightening the conceptual link between the mereological and the topological ingredients of MT?

Consider (C.6). As it turns out, its status depends significantly on the underlying axioms for 'P'. If our mereological theory is strong enough to warrant the existence of a fusion for any pair of connected entities, i.e., if it contains the relevant instance of (P.11$_\psi$) as an axiom or (as in GEM) as a theorem, then (C.6) itself is derivable as a theorem, since the fusion of x and y is sure to qualify as a z satisfying the consequent. If, however, our theory does not warrant all the relevant fusions, then (C.6) may still be regarded as a plausible addition to MT. Perhaps the fusion of any two connected entities turns out to be too large or gerrymandered to be acceptable, but we may still think of connection as sufficient for the existence of smaller fusions encompassing those portions of x and y that are sufficiently close to their common boundary. For example, with reference to Fig. 15.11, left, suppose we are only willing to acknowledge the existence of entities that are composed of at most two disconnected parts. Then the fusion of x and y is out, but a fusion of x_2 and y_2, as well as fusions of x and y_2 and of x_2 and y, would fit the bill. All of this speaks in favor of (C.6), though it may be argued that the existential import of this principle goes beyond the task of establishing a necessary conceptual link between P and C.

As for (C.7), the picture is different. Assuming this principle is virtually tantamount to excluding disconnected entities from the domain—not all of them, to be sure, but many of them. For example, if the underlying mereology is sufficiently weak, (C.7) is compatible with the existence of a disconnected composite such as c in Fig. 15.11, right: on the assumption that c has no further proper parts besides a_1, a_2, and a_3 (and parts thereof), the antecedent is false so (C.7) is vacuously satisfied. But consider a bikini, or a printed inscription consisting of separate letter tokens. As we have already noted, one need not buy into unrestricted composition to appreciate the dignity of such things. Yet their existence would be banned by (C.7). If x and y are the two main parts of a bikini, then the consequent of (C.7) is false even though the antecedent is made true by the bikini as a whole. So, on the face of it, this direction of Whitehead's biconditional is definitely too strong as a general bridging principle and there is no philosophically neutral reason to add it to MT. One can, however, consider weaker versions, to the effect that the consequent of the conditional must hold whenever the antecedent is made true by the right sort of entity:

$$\text{(C.7}_\phi\text{)}\quad \exists z(\phi z \wedge \mathsf{O}zx \wedge \mathsf{O}zy \wedge \forall w(\mathsf{P}wz \rightarrow (\mathsf{O}wx \vee \mathsf{O}wy))) \rightarrow \mathsf{C}xy$$

In particular, one can take 'ϕ' to express the property of being self-connected. After all, this was precisely the intended import of Whitehead's flawed definition. And we have seen that the flaw of the definition, in the *if* direction, was not conceptual but formal: it lied exclusively in the impossibility of expressing the relevant restriction in mereological terms. By making the restriction

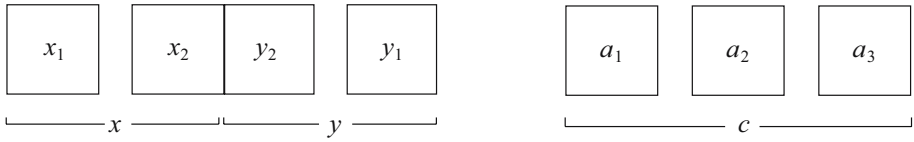


Figure 15.11. Left joining with partial fusions; right joining without connection.

explicit, (C.7$_\phi$) overcomes the difficulty and suggests itself as a natural bridging principle.

Surprisingly, it is not easy to express the property of self-connectedness even in the extended language of mereotopology. If the axioms on P are strong enough, we can follow the ordinary set-theoretic definition—something is self-connected if it doesn't consist of disconnected parts:

(75) $\mathsf{SC}x =_{df} \forall yz(\forall w(\mathsf{O}wx \leftrightarrow (\mathsf{O}wy \vee \mathsf{O}wz)) \rightarrow \mathsf{C}yz)$ Self-Conn.

In particular, in $MT + GEM$ this becomes:

(76) $\mathsf{SC}x \leftrightarrow \forall yz(x = y + z \rightarrow \mathsf{C}yz)$

This is a common definition in the literature, both among theorists subscribing to the converse monotonicity principle (C.4) (see e.g. Clarke 1981, Randell *et al.* 1992) and among theorists rejecting it (Tiles 1981, Varzi 1994, Smith 1996). However, if the axioms on P do not secure the necessary composition patterns, the definition is inadequate. For example, the object c in Fig. 15.11, right, is anything but self-connected, yet it (vacuously) satisfies the definiens of (75) unless we assume the existence of at least one sum consisting of a_1 and a_2, a_2 and a_3, or a_1 and a_3. In the finitary case, the difficulty could be met by relying on the notion of mediate connection: any two parts of a self-connected entity must be at least n-connected for some n. More generally:

(77) $\mathsf{PC}x =_{df} \forall yz(\mathsf{P}yx \wedge \mathsf{P}zx \rightarrow \mathsf{TC}yz)$ Path Connectedness

This, however, involves quantification over numbers (see (62)), which just confirms the expressivity limits in question. Moreover, (77) does not work in the infinitary case: the unit interval on the real line is connected, but we cannot account for this fact in terms of the relationships between the reals themselves; reference to subintervals is necessary, specifically reference to a subinterval and its relative complement. So, overall it appears that the notion of self-connectedness can be adequately grasped, via (75), only by theories that are at least as strong as MT + (P.6), the complementation principle, though this is at present an open question.

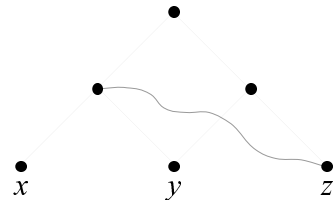


Figure 15.12. An implausible model of $MT + GEM$.

2.3.4 Fusions. We conclude this discussion of bridging principles by noting that even a theory as strong as $MT+GEM+$ any of the above axioms is incapable of capturing all the relevant links between mereology and topology. In particular, such a theory is consistent with the following implausible thesis (Tsai 2005, p. 137):

(78) $\exists z(\mathsf{C}z(x+y) \wedge \neg\mathsf{C}zx \wedge \neg\mathsf{C}zy)$

A model is given in Fig. 15.12, where the curve line indicates the relevant connection relationship (besides the obvious ones imposed by (C.3)).

Clearly, this is a sign that some additional bridging principle is on demand. The following option suggests itself:

(C.8) $z = \Sigma x\phi x \rightarrow \forall y(\mathsf{C}yz \rightarrow \exists x(\phi x \wedge \mathsf{C}yx))$ Fusion Connection

Whether this is enough to establish a good correlation between the mereological structure of composite objects and their topological behavior is a question that can hardly be addressed in general terms. The plausibility of (C.8), however, seems obvious. Since we have found good reasons to also accept (C.4) and (C.7$_{\mathsf{SC}}$), the theory resulting by adding these three principles to $MT + GEM$ suggests itself as the natural topological extension of GEM. (Recall that (C.6) is already provable in $MT + GEM$.) For future reference, we shall call this theory $GEMT$, for *General Extensional Mereotopology*.

2.4 Extensions and restrictions

As it turns out, $GEMT$ does not officially appear in the literature, mostly due to the limited study of such principles as (C.7$_{\mathsf{SC}}$) and (C.8). (The closest relatives are the axiomatic systems advocated by Smith 1996 and Casati and Varzi 1999.[11]) A good thing about this theory, as also about the core fragment afforded by $MT + GEM$, is that it makes it possible to supplement the

[11] In Casati and Varzi (1999), $GEMT$ is identified with $MT + GEM$. Smith's (1996) version is based on a primitive 'IP' for (possibly improper) interior parthood. See also Pianesi and Varzi (1996a, 1996b) for similar formulations based on the predicate 'B' and the operator 'c' defined below, respectively.

mereotopological predicates and operators discussed so far with a number of additional operators that mimic the standard operators of point-set topology. For example:

(79) $\mathsf{i}x =_{df} \Sigma z \forall y(\mathsf{C}zy \rightarrow \mathsf{O}xy)$ interior
(80) $\mathsf{e}x =_{df} \mathsf{i}(\sim x)$ exterior
(81) $\mathsf{c}x =_{df} \sim (\mathsf{e}x)$ closure
(82) $\mathsf{b}x =_{df} \sim (\mathsf{i}x + \mathsf{e}x)$ boundary

Like the mereological operators in (40)–(44), these operators are partially defined in view of the lack of a null entity that is part of everything. For instance, if x is a boundary, then it has no interior, and if x is the universal entity $\mathcal{U}$, it has no exterior. Even so, in $GEMT$ all of these operators are rather well-behaved. In particular, we can get closer to standard topological theories by explicitly adding the mereologized analogues of the standard Kuratowski (1922) axioms for topological closure:

(C.9) $\mathsf{P}x(\mathsf{c}x)$ Inclusion
(C.10) $\mathsf{c}(\mathsf{c}x) = \mathsf{c}x$ Idempotence
(C.11) $\mathsf{c}(x + y) = \mathsf{c}x + \mathsf{c}y$ Additivity

(These axioms are to be read as holding whenever c is defined for its arguments. Here and below we omit the relevant existential conditions to improve readability.) Indeed, (C.9) and (C.11) turn out to be provable in $MT + GEM$; see Tsai (2005, p. 141).

The possibility of supporting such developments is of course a good indication of the strength of $GEMT$. Philosophically, however, this strength may be regarded with suspicion, and several complaints have been raised in the literature.

2.4.1 The open/closed distinction. The main sort of complaint concerns the very notion of connection that the theory is meant to characterize. So far we have worked mostly with an intuitive notion in mind but obviously more can and must be said—and $GEMT$ says a lot. In particular, the Kuratowski extension of $GEMT$ ($KGEMT$ for short) yields a full account of the intended interpretation of 'C': two things are (externally) connected if and only if they share (only) a boundary, i.e., if and only if the closure of one overlaps the other, or vice versa:

(83) $\mathsf{C}xy \leftrightarrow (\mathsf{O}x(\mathsf{c}y) \vee \mathsf{O}(\mathsf{c}x)y)$
(84) $\mathsf{EC}xy \leftrightarrow (\mathsf{C}xy \wedge \neg\mathsf{C}(\mathsf{i}x)(\mathsf{i}y))$

Now, this shows in what sense the behavior of C in this theory closely approximates that of standard set-theoretic topological connection; just let 'x' and 'y'

range over sets of points and interpret 'O' as set intersection. On the other hand, one aspect in which ordinary point-set topology appears to conflict with common sense—an aspect that has been emphasized by authors interested in a mereotopological characterization of qualitative spatial reasoning, such as Randell *et al.* (1992) and Gotts *et al.* (1996)—is precisely the distinction between 'open' and 'closed' entities on which it rests, and which $GEMT$ preserves *holus bolus*:

(85) $\mathsf{OP}x =_{df} x = \mathsf{i}x$ Open

(86) $\mathsf{CL}x =_{df} x = \mathsf{c}x$ Closed

This distinction goes at least as far back as Bolzano (1851, §66ff). But already Brentano (1906, p. 146) regarded it as 'monstrous', and we have already seen that the sort of idealization it embodies does not sit well with the way we ordinarily speak. We may intuitively grasp the difference between an open and a closed interval on the real line—the objection goes—and we may even understand how this difference applies to ideal three-dimensional manifolds such as Euclidean space. But what does it mean to draw a similar distinction in the realm of concrete spatial entities, where the very notion of a boundary is the result of a conceptual idealization? What does it mean to say that some objects are closed and some are not, and that contact is only possible between objects of one type and objects of the other?

Besides, even if common sense and ordinary language were put aside, the open/closed distinction seems to yield genuine paradoxes as soon as we move from the realm of pure space to its worldly population: Consider (i) what happens when a solid body splits into two halves. Before the splitting the two halves were in contact, so we are to suppose that one was closed and the other open, at least in the relevant contact area. But then, after the splitting, only one of the two halves will have a complete boundary. This may not be 'monstrous', but it certainly seems implausible: the two halves—one should think—are perfectly indistinguishable. On the other hand, consider (ii) what happens when two bodies come into contact. We may imagine the same experiment performed twice. First we take an open cube and push it toward a closed cube until they touch. Then we do the same with two closed cubes. What reason can we offer to explain the fact that in the latter case the two cubes will *not* come into contact? As Zimmerman (1996a, p. 12) put it, what sort of 'repulsive forces' can be posited to explain such deferential behavior?

There is no quarrel that these are pressing questions. (For more examples, see Kline and Matheson 1987.) Nonetheless, there are various things one can say in reply to such worries. Concerning (i), for example, one could say that the paradox is grounded on a questionable model of what happens when a process of 'splitting' takes place (Varzi 1997). Topologically, this is no bloodstained business. Dissecting a solid body does not 'bring to light' (Adams 1984, p. 400)

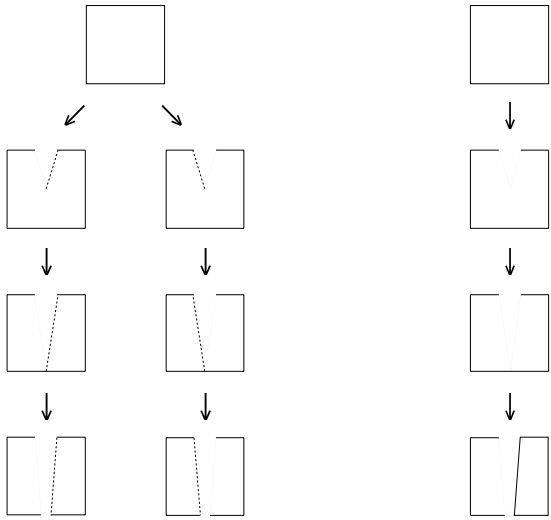


Figure 15.13. Wrong (left) and right (right) topological models of splitting.

a surface that was trapped inside and that must by necessity belong to one of two severed halves. Rather, the topological model is one of gradual deformation. Think of a splitting oil drop. The drop grows longer and, as it grows, the middle part shrinks and gets thinner and thinner. Eventually the right and left portions come apart and we have two drops, each with its own complete boundary. Ditto with any splitting object. A long, continuous process suddenly results in an abrupt topological change: there was one surface, one closed body, and now there are two (Fig. 15.13). Of course, one can still raise a question about the last point of separation: Where does this one point belong—to the left half or to the right half? However, this is just a sign of the magic that surrounds any sort of topological change, as when you drill a hole through an object. The instantaneous event of a sphere turning into a torus is just as magical, and this magic has nothing to do with the open/closed opposition. It simply reflects the fact that topological change marks one point at which common sense reaches the limits of its theoretical competence, and a complete assessment would require a step beyond pure part-whole theorizing. It would require a step into the territory of qualitative kinematics, for example (as in Davis 1993), if not an account in terms of the microscopic analysis of matter. (What is—physically—the 'last point of separation' involved in the splitting?)

This line of reply on behalf of $KGEMT$ applies to (ii) as well—the 'merging' puzzle (Casati and Varzi 1999). Surely the positing of repulsive forces to explain the peculiar behavior of the two closed cubes would be utterly and unagreeably *ad hoc*. But there are other possibilities. For instance, perhaps the two closed cubes *will* indeed come into contact. From the fact that two closed entities

cannot *be* in contact it does not follow that they cannot *come into* contact, just as from the fact that two parts are connected it does not follow that they cannot be separated. Only, the coming into contact (just as the separation) determines a true topological catastrophe: there is a breaking through the relevant boundary parts and the two objects become one. (Think also of the two drops of oil merging into each other.) The two processes are dual: merging is the reverse of splitting. And both involve a seemingly magic moment that runs afoul of the confines of extensional mereotopology and calls for a thorough kinematic account.

One could still press the objection here by noting that the puzzles admit of perfectly *static* variants, where the appeal to kinematics would be out of place. Consider the dilemma raised by Leonardo in his *Notebooks*: What is it that divides the atmosphere from the water? Is it air or is it water? Or consider Peirce's puzzle: What color is the line of demarcation between a black spot and its white background? (1893, p. 98). More generally, given any object, x, does the boundary belong to x or to its complement? Does it inherit the properties—e.g., color properties—of x or of $\sim x$? There is no kinematic story to tell here. But how can one answer without selecting one candidate at random?

Here one might reply that figure/ground considerations will help. According to Jackendoff (1987, Appendix B), for example, normally a figure owns its boundary—the background is topologically open. This may well be the right thing to say *vis-à-vis* Peirce's puzzle: the black spot is closed, so the line is black. But what is figure and what is ground when it comes Leonardo's case? We do talk about the surface *of* the water, not of the air. But what goes on at the seashore? Three things meet—water, air, soil; how can figure/ground considerations help in such contexts? Perhaps such dilemmas are not real. Galton (2003, pp. 167ff), for example, argues that they arise as an artifact of the modeling process: surely properties like color or material constitution only apply to extended bodies, so it wouldn't make sense to ask whether a boundary-like entity is air, water, or colored. There is, however, a less dismissive way to meet the challenge on behalf of $KGEMT$. For one may acknowledge that such dilemmas are real and yet insist on a friendly attitude towards the open/closed distinction. The actual ownership of a boundary—one might argue—is not an issue that a mereotopological theory must be able to settle. The theory only needs to explain what it means for two things to be connected. *Which* things are open and which are closed is a metephysical question that, plausibly enough, goes beyond the concerns of the theory. If the ocean is a closed body, then it can only touch the air if the latter is open. If it isn't, then it can only touch the air if the latter is closed. And if both water and air are open, they cannot truly touch, though they may touch a closed piece of land. That's all the theory says, and there is no reason to think that the theory is wrong just because it is difficult to classify actual things into open and closed.

One last problem is worth mentioning. Consider again the cutting of a solid object in half. We have said that this process does not bring to light a new surface. But, of course, we can *conceptualize* a new, potential surface right there where the cut would be. In fact, we can conceptualize as many boundaries as we like, even in the absence of any corresponding discontinuity or qualitative heterogeneity among the parts. Think of John's waist, the equator, the Mason-Dixon line between Maryland and Pennsylvania. As Smith (1995, 2001) has pointed out, such 'fiat' boundaries are a major ingredient of our picture of the world. Even the surfaces of ordinary objects, as we have seen, may involve a certain degree of fiat owing to the microscopic scatterdness of the underlying stuff. Yet *no fact of the matter* can support the ownership of boundaries such as these by one side rather than the other, hence there is no point in deferring to a metaphysical theory of the extended entities at issue. Isn't this enough to give rise to the demarcation puzzle?

Once again the answer is in the negative. Fiat boundaries are not physical boundaries *in potentia*. They are not the boundaries that would envelop the interior parts to which they are associated in case those parts were actually cut off. To think so would take us back to the wrong topological model illustrated in Fig. 15.13, top. On a better model, fiat boundaries are just *placeholders* for genuine physical boundaries and the demarcation puzzle need not, therefore, lead to ontological anxiety. We can say that a boundary of this sort stands for *two* boundaries, one for each side. Or we can say that in drawing it we leave the question of its belongingness (hence the open/closed distinction) indeterminate. We do so because it is a question of no practical relevance. But precisely for this reason the indeterminacy is innocuous: it is pragmatic, perhaps semantic, not ontological—just like the sort of indeterminacy that afflicts the vagueness of parthood (Sec. 1.5). (It is in this sense that the diagram in Fig. 15.9 is partly indeterminate: in saying that x_2 is externally connected to y, for example, we did not specify which of these two regions owns the boundary in the relevant contact area. Ditto for all other cases of external connection, as in Fig. 15.8, left.)

2.4.2 Connection by coincidence. All of this, of course, is subject to controversy. If the foregoing remarks are found compelling, then the strength of $KGEMT$ is vindicated and such theorems as (85) and (86) deliver a full and correct understanding of 'C'. If not, however, then $KGEMT$ will be deemed inadequate and the intended interpretation of 'C' remains unsettled. Are there any other options? We may distinguish two main alternatives, depending on whether or not a rejection of the open/closed distinction is taken to be compatible with a realist attitude towards the ontological status of boundaries.

The realist option finds its best expression in those theories that attempt to provide a detailed reconstruction of the view Brentano put forward in reaction

to Bolzano's 'monstrous doctrine', as in Chisholm (1984, 1993) and Smith (1997). According to this view, boundaries are genuine denizens of the world of spatial entities, but their lack of proper interior parts makes them peculiar in two important respects (Brentano 1976, part I). First, they can never exist except as belonging to entities of higher dimension. There are, in other words, no isolated points, lines, or surfaces, for boundaries are, in Chisholm's terms, *dependent* entities. Second, and more to the point, insofar as boundaries are not possessed of divisible bulk, they do not occupy any space and can therefore share the same location with other boundaries. They can *coincide*, and the topological relation of external connection is to be explained, not via the open/closed opposition, but in terms of genuine boundary coincidence. Thus, we can speak of *the* surface of an object. But this single surface is to be recognized as being made up of two parts, two perfectly coinciding boundaries bounding the object and its complement, respectively.

As is clear, a rigorous formulation of such theories is no straightforward business. For one thing, it is not immediately obvious how to formulate the dependence thesis, both because of the modal ramifications that a good theory of ontological dependence would require (see e.g. Correia 2005) and because the relevant notion of dimension is by itself hard to characterize mereotopologically (Chisholm 1984). Ignoring such complications, and assuming GEM, one can capture the gist of the thesis as follows (Smith 1996):

(C.12) $\mathsf{SC}x \wedge \mathsf{B}xy \rightarrow \exists z(\mathsf{SC}z \wedge \mathsf{BP}xz \wedge \neg\exists w\mathsf{B}zw)$ Dependence

where

(87) $\mathsf{B}xy =_{df} \mathsf{P}x(\mathsf{b}y)$ Boundary
(88) $\mathsf{BP}xy =_{df} \mathsf{B}xy \wedge \mathsf{P}xy$ Boundary Part

In other words, every self-connected boundary is part of some self-connected entity which it bounds and which is not itself a boundary. (The restriction to self-connected entities is to avoid that (C.12) be trivially satisfied by a scattered z containing x as an isolated proper part.) Without the full mereological support of GEM, however, things are significantly more complex, among other reasons because of the apparent elusiveness of the self-connectedness predicate 'SC' (Sec. 2.3.3.)

Secondly, and more to the point, a lot depends on how exactly one understands the relation of spatial coincidence invoked by such theories to explain the phenomenon of (external) connection. Chisholm and Smith treat it as an undefined primitive, suitably axiomatized so as to guarantee that coinciding entities have coinciding parts. (See also Smith and Varzi 2000 for a similar treatment of the relation of coincidence between *fiat* boundaries.) Alternatively, one can embed the theory of coincidence into a theory of spatial location

broadly construed: to say that things coincide is to say that they literally share the same location. Clearly, the choice between these two options is not just a matter of taste. Treating coincidence as a primitive is in principle compatible with different metaphysical conceptions of the nature of space, whereas the second option is best understood within the framework of a substantivalist (Newtonian) conception, i.e., a conception according to which space is an entity in its own right. In any event, it is apparent that both options yield theories that are not strictly mereotopological, since a third primitive—coincidence or location, respectively—needs to be brought into the picture to provide a full account of the connection relation. We shall not go into the details of the first option here, but we shall have more to say about the second option in Sec. 3. For the moment, let us just observe that construing coincidence explicitly in terms of spatial co-location amounts to a partial reduction of topology to mereology: connection between entities of a kind (space occupiers) is reduced to overlap between entities of a different kind (their spatial receptacles), as per the following principle:

(89) x is connected to y if and only if the region occupied by x overlaps the region occupied by y.

This is by itself interesting, though we are obviously left with the task of providing an account of the topology of space as such. And if the account is to match the strength of a theory such as $KGEMT$, then the open/closed distinction will at least be partially preserved. It will be obliterated from the Brentanian realm of space occupiers, but space itself would be Bolzanian. (Compare the initial worry: we *can* grasp how the distinction applies to ideal manifolds such as the real line or Euclidean space; it is when it comes to the realm of ordinary objects that their classification into 'open' and 'closed' is problematic.)

2.4.3 Omitting boundaries. The alternative route is to avoid the puzzles raised by the open/closed distinction by dismissing boundary talk altogether. This is the anti-realist option.

Philosophically, this route is often motivated on its own grounds, for instance because of the dubious ontological status of boundaries *vis-à-vis* the microscopic analysis of the physical world (Stroll 1988), or because of their suspect nature *qua* lower-dimensional entities (Zimmerman 1996b). In the context of formal theories, however, the main motivation for doing away with boundaries is precisely the rejection of the open/closed distinction *vis-à-vis* common sense. To use an example from Gotts *et al.* (1996, p. 57), Fig. 15.14 depicts a disc with and without its boundary, and with just part of its boundary. Of course, the depiction has to show the boundary as having some finite thickness, which strictly speaking it does not possess. But this is the very point that appears counter to common sense: all three discs, if superimposed, would cover exactly the same

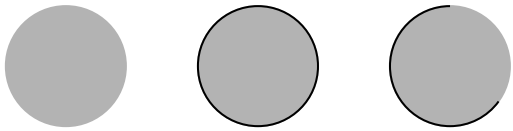


Figure 15.14. Open, closed, and semi-closed discs with the same area.

area; yet the second disc includes unextended parts that the others do not, while the third includes some that the first does not. Such discriminations—it is argued—are not warranted.

There are radical as well as moderate variants of this view. The radical variants are represented by those theories that follow Whitehead (1929) in doing away with *all* boundaries. This amounts to assuming the boundarylessness axiom (C.5) in its full strength: everything has interior proper parts. The moderate variants, by contrast, only assume some weaker version of the axiom in which the variable is suitably restricted so as to range over entities of a certain sort:

(C.5$_\phi$) $\phi x \rightarrow \exists z \mathsf{IPP} yx$ Restricted Boundarylessness

In particular, relative to our present concerns it is natural to construe 'ϕ' as a distinguished property of all concrete spatial bodies. This would allow for the possibility that space as such include points and other boundary-like elements, which means that the open/closed distinction would be partially preserved. But as we have just seen, restricting the distinction in this way may be enough to bypass the intuitive puzzles that it raises, so this may well be a good compromise. For example, Cartwright (1975) holds that concrete spatial bodies are the material content of (regular) open regions of space, connection relations between the former being explained in terms of overlap relations between the closures of the latter:

(90) x is connected to y if and only if the closure of the region occupied by x overlaps the closure of the region occupied by y.

As a mereotopological theory, this is of course another hybrid—just as the theory behind (89)—for it requires an explicit treatment of locative relations. Still, there is no question that (90) allows for a systematic boundary-free account of the mereotopology of concrete spatial bodies. (The real challenge, rather, is to justify the claim that only *some* regions are receptacles, e.g., only open regular regions; see Hudson 2002).

Let us focus on the radical variants. We have seen that positing (C.5) is a necessary move for any reductive mereotopology based on the converse monotonicity axiom (C.4), and it is a fact that most theories that accept one axiom

accept the other as well. But let us put that aside for a moment and let us just focus on (C.5). Where X is any theory including MT, let $\bar{B}X$ be the corresponding *boundaryless* extension obtained by adding this axiom. What sort of mereotopology do we get?

As it turns out, the number of options is significantly constrained, both mereologically and topologically. For example, surely X cannot be atomistic, since (C.5) implies the atomlessness axiom (P.7). So any model of $\bar{B}X$ is perforce infinitary. And surely the interaction between compositional and decompositional principles will have to be carefully re-examined. In particular, it is easy to verify that in $\bar{B}MT$ the weak supplementation principle (P.4) is incompatible with the unrestricted fusion axiom (P.13) and, more generally, with any version of the strong fusion axiom (P.13$_\xi$) in which the condition 'ξ' is satisfied by all interior and tangential proper parts of any given thing. For suppose we allow for such fusions. Then every entity would have an interior as well as a closure, and the following would hold:

(91) $\mathsf{PP}(\mathsf{i}x)(\mathsf{c}x)$.

By (P.4), this would imply

(92) $\exists z(\mathsf{P}z(\mathsf{c}x) \wedge \neg\mathsf{O}z(ix))$,

which in turn would imply

(93) $\exists z\mathsf{P}z(\mathsf{b}x)$,

contradicting (C.5). Thus, $\bar{B}MT$ + (P.4) + (P.13) is inconsistent, as is any theory $\bar{B}X$ including (P.4) along with (P.13$_\xi$) with 'ξ' as indicated. This is not surprising, since the whole point of going boundary-free is, in the present context, to avoid the open/closed distinction, hence the distinction between interiors and closures reflected in (91). However, this means that (C.5) prevents the formulation of any reasonably strong theory unless we are willing to give up weak supplementation, and this may certainly be regarded as a major drawback of the approach.

In fact, one may consider both options here. One may (i) regard the compositional weakness of the theory as a necessary price to pay to preserve mereological supplementation and avoid the topological conundrums surrounding the open/closed distinction. But one may also (ii) go for a stronger theory with generalized or even unrestricted fusions, dismissing the conundrums precisely by forgoing supplementation. After all, it could be argued that the open/closed distinction is problematic *only* insofar as it is cashed out in terms of a discrimination between entities that do and entities that do not possess their boundaries, and in the absence of boundaries the discrimination dissolves. In the literature the first option is more widespread, its closest representative being the greatly influential

Region Connection Calculus (RCC) originated with Randell *et al.* (1992)—a reductive extension of $\bar{B}MT+$ (P.5) with binary sums and complements. But the second option, which is closer to Whitehead's original approach, is also well represented, as evidenced by Clarke (1981), Biacino and Gerla (1991), and Asher and Vieu (1995) *inter alia* (the first two admitting unrestricted fusion, the latter admitting fusions of interior parts). Indeed, we have seen in Sec. 1.4.2 that the idea of restricting mereological fusion in order to avoid undesired entities is by itself suspicious. If we agree with the thought that a fusion is nothing over and above the things that compose it, then the intuitive problems raised by the open/closed distinction are hardly solved by eschewing formal commitment to such things as interiors and closure, i.e., fusions of interior and of tangential parts. For such parts are *all* there already (and, of course, one needs IPP and TPP to be distinct in order to state (C.5) in the first place). In this sense, option (ii) might be regarded as preferable on philosophical grounds, though the failure of supplementation would remain a hindrance.

Unfortunately, all of these theories include the monotonicity axiom as well as its converse, (C.4), i.e., they are all of the reductive sort, which makes it difficult to assess their relative pros and cons *vis-à-vis* the two options in question. In fact, not only do such theories include (C.4) and (C.5); they also rest on a *sui generis* characterization of the fusion operator in which C takes over the role of O, which makes it difficult to compare them to $KGEMT$. For example, in Clarke's theory, which goes as far as to include analogues of the Kuratowski axioms, the unrestricted fusion principle does not equal (P.13) but, rather, the following schema:

(C.13) $\exists w\phi w \rightarrow \exists z\forall w(\mathsf{C}wz \leftrightarrow \exists v(\phi v \wedge \mathsf{C}wv))$ Topological Fusion

This leads to a correspondingly *sui generis* fusion operator:

(94) $\Sigma^* x\phi x =_{df} \imath z\forall w(\mathsf{C}wz \leftrightarrow \exists v(\phi v \wedge \mathsf{C}wv))$ fusion*

(Recall that RMT treats C as extensional: see (73).) And it is easily checked that Σ^* does not coincide with the operator Σ defined in (39): if parthood reduces to enclosure, then the interior of a closed entity x qualifies as the fusion of x's proper parts, but not as their fusion*. (This is because in the absence of boundaries the interior is sure to overlap any y that x overlaps, and vice versa, though it will to be disconnected from any z to which x is *externally* connected; see Fig. 15.15). This amendment is plausible enough. But it means that the mereotopological operators defined in (40)–(44) and in (79)–(82) must be revised accordingly, and at the moment there is no systematic comparison between the behavior of such revised operators in a boundary-free theory and the behavior of the original operators in a boundary-tolerant theory. Just to give an example, note that re-defining '$\sim$' in terms of Σ^*:

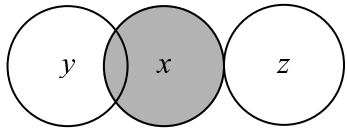


Figure 15.15. The interior of a closed disc x is a fusion, but not a fusion*, of its parts.

(95) $\sim^* x =_{df} \Sigma^* z \mathsf{D} z x$ complement*

implies that nothing is connected to its own complement and, therefore, that the universe is bound to be disconnected. Of course we can rely on a different notion of complement (as suggested in Randell *et al.* 1992, p. 168), or one can change the definition of self-connectedness in such a way as to avoid at least the latter consequence (as suggested in Clarke 1985, p. 69). But this is playing with definitions. The 'old' notions, when revisited in terms of (94), continue to make good sense no matter how we change the official nomenclature, so we can hardly leave it at that.

Perhaps the best way to assess the strengths and weaknesses of these theories is to note that the departure from the ordinary conception of fusion affects the very distinction between open and closed entities. Consider the following variants of (79) and (85):

(96) $\mathsf{i}^* x =_{df} \Sigma^* z \forall y (\mathsf{C} z y \rightarrow \mathsf{O} x y)$ interior*

(97) $\mathsf{OP}^* x =_{df} x = \mathsf{i}^* x$ Open*

It is easily checked that any theory at least as strong as $\bar{B}RMT$ + (C.13) has the following theorem:

(98) $\mathsf{OP}^* x \rightarrow \neg \mathsf{EC} x y.$

Thus, open* entities never touch anything. It is only closed* and semi-closed* entities (defined similarly), that can touch something without sharing any parts.[12] And if the open/closed distinction is replaced by the open*/closed* distinction, then the intuitive import of the relevant misgivings is up for grabs, and the choice between a conservative attitude towards mereological supplementation (option (i)) and a liberal attitude towards mereological composition (option (ii)) calls for independent thinking. We are no longer dealing with a partitioning of the domain into entities that do and entities that do not possess a boundary;

[12]Actually, Clarke's definition of the closure operator does not exactly parallel (81). In our notation it reads as follows:

$\mathsf{c}^* x =_{df} \Sigma^* z \neg \mathsf{C} z (\mathsf{i}^* x).$

However, this peculiarity does not affect the main point made in the text.

we are dealing with a partitioning into entities that do and entities that do not connect externally.

Be that as it may, all of this suggests that a thorough comparison between these two strategies for construing boundaryless mereotopologies is a challenging task (Cohn and Varzi 2003). What is clear is that the strategies are mutually incompatible in spite of their common motivation and this, in all fairness to $KGEMT$, is disturbing. No matter how one feels about subtracting or adding elements to the domain, there is something puzzling in the thought that a topological 'monstrosity' should by cured through mereological surgery. Indeed, philosophically this puzzling feature is especially striking when it comes to explaining the intended *interpretation* of these theories. As it turns out, both can be modeled on a domain with a standard point-set topology, interpreting 'C' as in (99) for type-(i) theories, and as in (100) for type-(ii) theories:

(99) x is connected to y if and only if the closure of x and the closure of y have a point in common.

(100) x is connected to y if and only if x and y have a point in common.

(See e.g. Randell *et al.* 1992, p. 167, Gotts 1996b and Pratt and Schoop 2000 *vs.* Clarke 1981, p. 205, Biacino and Gerla 1991, and Asher and Vieu 1995, respectively). There is, of course, nothing wrong with this sort of models when it comes to proving the consistency or even the completeness of such theories. And there would be nothing wrong with (99) and (100) as genuine models if we were dealing with boundaryless theories of the moderate sort, as seen above. It is disturbing, however, that one can hardly do any better when it is comes to theories that are meant to be radically eliminativist—when it comes to explaining how contact relations may obtain in a world that is truly lacking the topological glue provided by points, lines, and surfaces even in the realm of pure space. (Whether one *can* do better is an open question. For example, with reference to type-(i) theories, Bennett 1996a suggests that RCC can be interpreted by encoding it into the bimodal propositional modal logic $\mathbf{S4}_u$, though the encoding is imperfect, as shown in Aiello 2000, and its natural canonical model is itself topological, as evidenced in Renz 1998 and Nutt 1999. Likewise, Stell and Worboys 1997, Stell 2000, and Düntsch *et al.* 2001 provide algebraic interpretations of RCC that dispense with any reference to point-based topologies, but the ontological transparency of such interpretations is itself a delicate matter.)

2.5 Expressivity and ontology

Let us conclude this philosophical excursus on topology with some general considerations concerning the delicate interplay between the expressive power of a theory and its ontological presuppositions. Regardless of whether we rely

on the full strength of $KGEMT$ or on theories of weaker import, we have seen that the move from mereology to mereotopology represents an important step towards the formulation of an adequate model of our spatial competence. The mereological distinction between a whole and its parts is crucial, but so is the distinction between interior and tangential proper parts, or the distinction between a connected whole and a scattered one, and such distinctions are intrinsically topological. This is not to say that mereotopological concepts exhaust the picture; geometric and morphological considerations also play a significant role when it comes to practical matters. But there is no question that a great deal of our spatial competence is grounded on our capacity to 'parse' the world in terms of parthood and connection relationships. The interesting question, rather, is whether such relationships can be fully captured by the behavior of the binary predicates 'P' and 'C' when characterized by means of formal principles of the sort that we have been discussing up to now, and to what extent the answer depends on one's specific views when it comes to matters of ontology. Here are some indicative examples.

2.5.1 Modes of connection. So far we have followed the familiar course of explaining connection in general terms, i.e., irrespective of the size (dimension) of the relevant contact area. In introducing that notion, however, we have mentioned the possibility of distinguishing connection ties of different strength—e.g., ties involving a single point of contact (as between Colorado and Arizona), an extended portion of a common boundary (France and Germany), or an entire boundary (Vatican and Italy). Even without bringing in boundaries, one may want to draw such distinctions to fully grasp, for example, the difference between a whole consisting of two spheres that barely touch from the whole consisting of two halves of a single sphere: both wholes are self-connected, but the second is surely more firmly connected than the first. And these distinctions have ramifications. For example, since tangential parthood is defined in terms of external connection, we may want to distinguish those proper parts that barely touch the exterior from those that firmly touch it—and so on. Moreover, the number of distinctions grows with the dimensionality of the entities we consider. Fig. 15.16 illustrates the four main patterns of external connection (no overlap) that can be distinguished in 2D space. But in 3D we might want to further distinguish, for example, two cubes barely touching at a vertex, two cubes barely touching along an edge, two cubes touching along a face, and so on. Now, can all such distinctions be expressed in terms of the mereotopological primitives 'P' and 'C' (or just 'C', if one goes reductive)?

As it turns out, within a sufficiently rich mereotopological theory such as $KGEMT$ the answer is in the affirmative. To illustrate, with reference to Fig. 15.16 the difference between the first two cases can be explained as follows. In both cases, x and y are connected; but whereas in the first case this simply

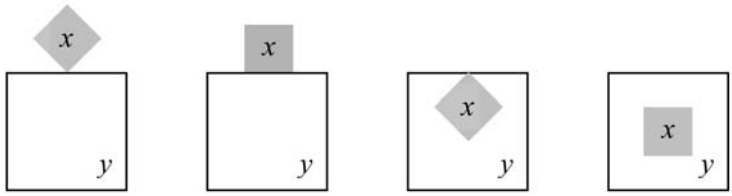


Figure 15.16. Four patterns of external connection, from weakest to strongest.

means that one can go from x to y without ever going through the exterior of the connected sum $x + y$, in the second case it is also possible to go from x to y without ever leaving the interior of $x + y$. More precisely, let us distinguish between the connectedness of a whole ('SC') and the connectedness of its interior, and let us say that an object is firmly self-connected just in case the latter condition holds:

(101) $\mathsf{FSC}x =_{df} \mathsf{SC}x \wedge \mathsf{SCi}x$ Firm Self-Connectedness

Then we can say that two entities are firmly connected when they have parts that add up to a firmly self-connected sum:

(102) $\mathsf{FC}xy =_{df} \exists wz(\mathsf{P}wx \wedge \mathsf{P}zy \wedge \mathsf{FSC}(w + z))$ Firm Connection

This captures the difference between the second case of Fig. 15.16, where the relevant connection relationship is firm, and the first, where it isn't. The stronger connection patterns corresponding to the third and fourth cases can then be defined by reference to the complement of $x + y$:

(103) $\mathsf{CC}xy =_{df} \mathsf{FC}xy \wedge \neg\mathsf{FC}x(\sim (x + y))$ Complete Connection
(104) $\mathsf{PC}xy =_{df} \mathsf{FC}xy \wedge \neg\mathsf{C}x(\sim (x + y))$ Perfect Connection

(Strictly speaking, (103) and (104) would call for refinements, owing to the possibility that x and y have internal holes; see Cohn and Varzi 2003 for a more general picture.) On this basis, the generalization to spaces of higher dimensionality is not difficult. For instance, in 3D space the difference between two entities touching at a point and two entities touching along an edge can be described by further distinguishing two patterns of non-firm connection, depending on whether the common boundary is atomic or a self-connected composite (a line segment, or curve).

Now, all of this is easy in $KGEMT$. When it comes to weaker theories, however, things get more difficult. For the definitions above involve the self-connectedness predicate along with the interior, complement, and fusion operators, all of which may be absent in a theory deprived of the necessary compositional strength. For instance, in a boundaryless theory with no open/closed

(or open*/closed*) distinction, we can supply for the lack of the interior operator by redefining firm self-connectedness as follows (Bennett 1996b, p. 345):

(101′) $\mathsf{FSC}x =_{df} \forall y(\mathsf{IPP}yx \rightarrow \exists z(\mathsf{IPP}zx \wedge \mathsf{P}yz \wedge \mathsf{SC}z))$

This makes it possible to go ahead with definitions (102)–(104) and capture the relevant distinctions in the 2D case. (See also Borgo *et al.* 1996 for a boundaryless theory in which FC is treated as primitive.) Yet it is not clear how one can capture the further distinctions available in 3D and higher-dimensional spaces without appealing explicitly to the dimensionality of the relevant boundaries or to special assumptions concerning the structure of space (see Gotts 1994a). And of course things get worse in a theory that lacks the complementation principle (P.6), for in that case, as we have seen, the notion of self-connectedness is already problematic. In short, the expressive power of a theory depends crucially on the underlying ontology, which in turn is reflected in the strength of the relevant compositional and decompositional principles.

Similar considerations apply to further conceptual distinctions that may be deemed relevant in the context of spatial reasoning. Consider again the common-sense notion of contact exemplified by such statements as (55): the table is touching the wall. We have said that this notion is not topological but metric. This is actually true in $KGEMT$, assuming that both entities—table and wall—are treated alike, i.e., as both closed or both open. For $KGEMT$ has the following two theorems:

(105) $\mathsf{EC}xy \rightarrow (\mathsf{CL}x \rightarrow \neg\mathsf{CL}y)$
(106) $\mathsf{EC}xy \rightarrow (\mathsf{OP}x \rightarrow \neg\mathsf{OP}y)$

By contrast, in a boundaryless theory the picture is different. For example, insofar as RCC admits of models satisfying (99), it can treat the table and the wall as genuinely connected as long as their closures overlap. This may be at odds with physics, but it captures the common-sense intuition. (Ditto for a moderate variant such as Cartwright's—see again (90).) So which of these accounts is better depends on the seriousness with which we handle the spatial ontology of common sense. On the other hand, none of this should be taken to imply that $KGEMT$ lacks the resources to account for the loose notion of connection countenanced by common sense. Suppose we understand this notion in the following sense: the table is touching the wall insofar as nothing *can* be squeezed between them. The metric flavor of this notion lies in its modal ingredient: to say that nothing can be squeezed between two objects is to say that they are 'vanishingly close' to each other—that their relative distance is arbitrarily small. We can, however, define a predicate that captures this ingredient in mereotopological terms: we can say that two objects are at least loosely connected in the relevant sense when one is connected to the closure of every open neighborhood of the other:

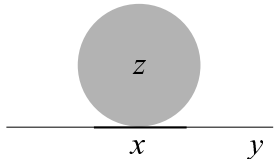


Figure 15.17. An interior tangential part?

(107) $\mathsf{LC}xy =_{df} \forall z(\mathsf{OP}z \wedge \mathsf{P}yz \rightarrow \mathsf{C}x(\mathsf{c}z))$ Loose Connection

(See Asher and Vieu 1995 for a similar definition.) This relation captures the intuition that nothing can lie between two entities that touch, even when those entities are closed. And surely enough the definition is consistent with $KGEMT$. So it is not that $KGEMT$ lacks the conceptual resources to do justice to common sense. It is, rather, that the relation of loose connection is bound to be empty in those models of $KGEMT$ where the open entities form a dense ordering, i.e., where the following holds:

(C.14) $\mathsf{OP}x \wedge \mathsf{OP}y \wedge \mathsf{PP}xy \rightarrow \exists z(\mathsf{OP}z \wedge \mathsf{PP}yz \wedge \mathsf{PP}zy)$ Open Density

And whether all models should satisfy this axiom is, on the face of it, a question about the ontological make-up of the world. (One can argue that as long as the open/closed distinction holds, common sense only requires a denial of (C.14) when the variables are restricted to the range of ordinary entities, as opposed to their spatial receptacles. For a full account of the mereotopology of discrete space, see Galton 1999 and 2000, §2.6, and the generalizations in Li and Ying 2004.)

2.5.2 Dimensionality. Consider a second example. We have seen that in a boundaryless theory everything must have interior proper parts, as per (C.5). However, the notion of an interior part is itself, in a way, a relative one, depending on the dimensionality of the space we are considering (Galton 2004). In 1D space, the middle portion of a line segment y, or any portion x that does not extend to y's extremities, would qualify as an interior part of y, so y itself would satisfy the axiom. As soon as y is embedded in 2D (or higher) space, however, all of its parts would be tangential, as they can all be connected to things to which y itself would be externally connected—e.g. a disc z (Fig. 15.17). Thus, in such higher spaces y would not satisfy (C.5).

Now, this is not by itself a disturbing fact if we take boundaryless theories to reflect a general intuition to the effect that all entities in the domain are of equal dimensionality. On the other hand, this very intuition is philosophically problematic. It is puzzling that whether something exists—a line segment, for instance—should depend on the dimension of space, for one may want to

declare one's ontological commitments while remaining neutral with respect to the difficult question of the dimensionality of space. Indeed, this becomes a necessity if one does not have the resources to address such a question in the first place. In ordinary point-set topology, one can say that a domain is of dimension $< n$ if and only if every open cover $O_1, \ldots, O_k$ can be refined to a closed cover $C_1, \ldots, C_k$ such that every point occurs in at most $n+1$ of the C_is. This characterization is not available in mereotopology unless one goes second-order, and surely it cannot be mimicked in a boundaryless theory that eschews the open/closed distinction. Gotts (1994b) and Bennett (1996b) suggest a way to bypass such difficulties by means of a different characterization, which only requires P (or rather: E) to be closed under the operations of complementation and binary sum, but the adequacy of such proposals is an open question. (By contrast, Galton 1996 shows that an adequate characterization is available in a boundary-based theory such as $KGEMT$, or weaker variants thereof; see also the layered mereotopologies of Donnelly and Smith 2003 and Donnelly 2004). So, again, we see here how the expressiveness of the theory depends crucially on the ontology it countenances, which in turn may be hard to specify within the theory itself.

2.5.3 Counting the holes. Finally, consider an example that lies somewhere between the previous ones—the notion of a simply connected whole, i.e., intuitively, a whole with no holes. Topologically, as also from the perspective of common sense, this notion is just as significant as the notion of a self-connected whole. Just as there is a big difference between an apple and a bikini, there is a big difference between an apple and a donut. Indeed, topology is often defined, intuitively, as a sort of rubber-sheet geometry that focuses precisely on these two differences, ignoring shape, size, and all sorts of other spatial properties that concern geometry broadly understood. Now, we have seen that self-connectedness is easily defined in mereotopological terms, though the adequacy of the definition presupposes the complementation principle (P.6) and is not, therefore, entirely neutral from an ontological standpoint. What about simple connectedness? More generally, the definition of self-connectedness can be extended so as to classify every object in terms of the (maximum) number of self-connected parts of which it consists—what is sometime called its 'separation number'. This can be done by defining a corresponding sequence of predicates, one for each positive integer, as follows:[13]

(108) $\mathsf{SN}_1 x =_{df} \mathsf{SC} x$ Separation$_1$

(109) $\mathsf{SN}_{n+1} x =_{df} \exists yz(x = y + z \wedge \neg \mathsf{C} yz \wedge \mathsf{SC} y \wedge \mathsf{SN}_n z)$ Sep.$_{n+1}$

[13]Here '$n + 1$' indicates arithmetical addition, as opposed to mereological summation.

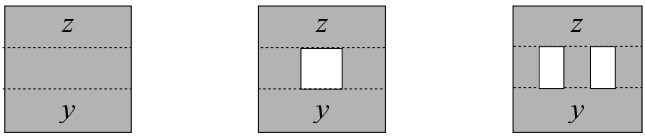


Figure 15.18. Simple connectedness and genus classification in 2D.

Can we also provide a mereotopological characterization of the genus of an object, so as to classify it in terms of the number of holes it has? As it turns out, the answer to this question is in the affirmative, again provided that we assume the complementation principle (P.6), but this affirmative answer has interesting ontological ramifications.

Here is how the basic account goes (Gotts 1994a). Let us say that something has dissectivity n (n a positive integer) just in case it is self-connected and can be decomposed into $n+2$ self-connected, disjoint parts with the following property: two of them, y and z, are not connected, whereas all the others are connected to both of them but disconnected from one another. Formally, this amounts to the requirement that there be two disconnected parts y and z that are connected by a remainder whose separation number is n ('connected by' in the sense of (59)):

(110) $\mathsf{DS}_n x =_{df} \mathsf{SC}x \wedge \exists yzw(x - (y+z) = w$
$\wedge\ \mathsf{BC}yzw \wedge \neg\mathsf{C}yz \wedge \mathsf{SN}_n w)$ Dissectivity$_n$

Then we can say that something is simply connected just in case its maximum dissectivity equals 1:

(111) $\mathsf{SSC}x =_{df} \mathsf{DS}_1 x \wedge \neg\mathsf{DS}_2 x$ Simple Connectedness

With reference to Fig. 15.18, for example, only the left pattern is simply connected, for the others have higher dissectivity numbers. This is how it should be, and it can be checked that the definition would yield the right classification also with reference to objects of different dimensionality. A donut has dissectivity 2, so it is not simply connected; a solid ball is. Indeed, we can now use (110) also to provide a mereotopological characterization of the genus of an object. Something is of genus n, i.e., has n holes ($n \geq 0$), if and only if its maximum dissectivity is $n+1$:

(112) $\mathsf{G}_n x =_{df} \mathsf{DS}_{n+1} x \wedge \neg\mathsf{DS}_{n+2} x$ Genus$_n$

Again, Fig. 15.18 illustrates the definition in the 2D case, but it can be checked that (112) yields the correct classification also for objects of higher dimensionality: a solid ball has genus 0, a donut has genus 1, a pretzel has genus 2, and so on.

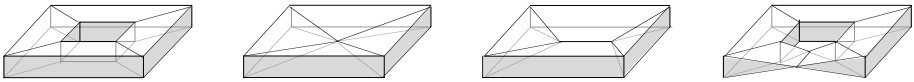


Figure 15.19. A donut, a punctured block, and other deviant solids of the same genus.

So this is the basic account, which fully answers the questions above: provided we work with a mereotopology that is closed under complementation and binary sums, we can define simple connectedness and, more generally, classify any object in terms of the number of holes it has. There are, however, two additional questions one may ask at this point. The first is whether this basic account can be refined so as to do justice to further distinctions that could be drawn in view of the dimensionality issues discussed above. We may, for example, want to tell a genuine donut from the deviant cases in Fig. 15.19, all of which have the same genus. And here, as one might expect, the answer depends more heavily on the strength of the theory. In $KGEMT$ we can go quite far; in weaker theories we may not, as we may not be able to distinguish the various kinds of non-firm connection that are needed to operate the relevant discriminations. For example, Gotts (1994b, 1996a) has shown that the issue can eventually be settled within the boundaryless framework of RCC, but this result rests on various assumptions on the topology and dimensionality of the entities in the domain that cannot themselves be expressed in the language of the theory. More importantly, it rests on the interpretation of 'C' given in (99): two bodies are connected if and only if their closures have a point in common. We have seen that such an interpretation is of dubious legitimacy in a boundaryless ontology (as Gotts himself laments in 1996b). So, once again, we reach a point where the expressive power of a theory depends crucially on the ontological commitments that one is willing to make.

The second question is whether the basic account can be refined so as to do justice to further distinctions that could be drawn in view of the various ways—and there are many—in which an object can be perforated. And here it appears that even a strong, boundary-based theory such as $KGEMT$ may show its limits. In fact, there is a sense in which the limits in question are not just the limits of the *mereo*topological approach of which the theory is expression; they are the limits of topology as a general theory of space. Let us focus on the 3D case.

For one thing, we have been speaking of 'holes' in the sense of perforations, but we may also want to classify a self-connected object in terms of the number of its internal 'cavities'. To some extent this is easy: on the assumption that the universe $\mathcal{U}$ is self-connected, it is sufficient to identify the number of internal cavities with $n-1$, where n is the separation number of the object's complement. In 2D space, this coincides with the genus of the object—there is no difference

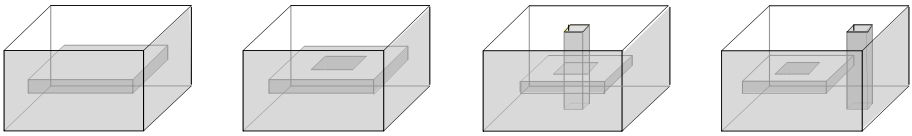


Figure 15.20. Cavities, donut-cavities, and donuts with donut-cavities.

between a 2D perforation and a 2D cavity. In 3D the numbers may diverge: a solid donut has genus 1 but 0 cavities, since its complement is self-connected. Dropping the assumption on $\mathcal{U}$, we can express this as follows:

$$\text{(113)}\quad \mathsf{IC}_n x =_{df} \mathsf{SC}x \wedge \forall y(\mathsf{SC}y \wedge \mathsf{IPP}xy \rightarrow \mathsf{SN}_{n+1}(y - x)) \qquad \text{Cavity}_n$$

This definition works for every object in any dimension (except, of course, for $\mathcal{U}$). But this is just the beginning. A cavity may come in different forms. It may be a solid cavity, so to say, but it may itself be donut-shaped. It may also have the shape of an irregular donut of the sort illustrated in Fig. 15.19. Or it may be 'knotted' in various fashions—as a trefoil knot, for instance, or a granny knot. Clearly such distinctions are not covered by (113). Moreover, consider an object with two donut-shaped cavities. The cavities may lie next to each other, so to say, or they may be interlocked like the rings of a necklace. Or consider an object with a perforation—a donut—which also has an internal, donut-shaped cavity. The perforation may go through the 'hole' in the cavity or it may lie next to it. All of these and many others are distinctions that are easily described in words just as they can easily be depicted (Fig. 15.20), and reflect significant differences in the spatial structure of the objects in question. It is far from clear, however, whether one can capture them in mereotopological terms.

Secondly, a perforation may come in different forms, too. It can be straight or it can be knotted, and the knot may or may not wrap around another perforation, just as it may or may not go through the 'hole' of an internal, donut-shaped cavity. It can also branch in the middle, so as to have more than two openings. Indeed, it can branch in many different fashions, as it can 'merge' in various ways with internal cavities of various kinds. Again, all of these possibilities reflect significant distinctions that are easily described in words and can easily be depicted, but it is far from clear whether one can give a proper characterization in mereotopological terms—even with the full strength of $KGEMT$. In some cases it is not even clear to what extent such a characterization should just parallel the standard topological account of such patterns. Standardly, for example, a block with two parallel, straight perforations is equivalent to a block with a single, Y-shaped perforation—both have genus 2 (and can be transformed into each other by mere elastic deformation). This much *can* be said in mereotopological terms, using (112) above. But here is where standard topological considerations

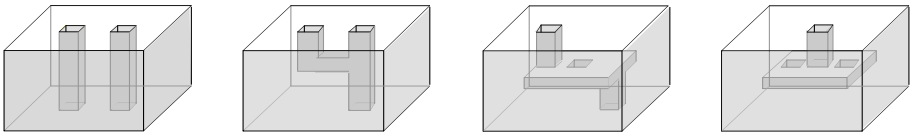


Figure 15.21. Same genus, different holes.

might be regarded as inadequate for a good description of the spatial structure of ordinary objects, and of the intuitions underlying our spatial reasoning broadly understood. The topological equivalence between such patterns—and between such patterns and many others; see Fig. 15.21—appears to deliver a partial account of the relevant spatial structures, for the genus of an object only captures the *intrinsic* topology of the object, not the way it relates to the environment. To get a better picture it seems necessary to keep an eye on the holes, not just on the object. And this is obvious from the fact that in describing such patterns we tend to do so by describing the mereotopology of the holes and the way *they* relate to each other; we do not describe the objects themselves. We tend to treat holes as objects in their own right, as 'negative objects' about which we can say exactly the same sorts of thing we say about ordinary, 'positive' objects. And we count both sorts of objects in the same way: we count *two* straight perforations and *one* Y-shaped perforation.

If this is correct, then there are two things one can say. One can say (and accept) that the limits of mereotopology *vis-à-vis* such fine-grained distinctions are just the limits of topology, *mutatis mutandis*. Or one can say that the limits in question reflect precise ontological assumptions concerning the domain of application of the theory, specifically a dismissive attitude towards the ontological status of holes. This is not to say that holes are left out of the picture. Surely any theory with unrestricted fusions has room for such things, for mereologically speaking a hole is nothing but part of the object's complement. Rather, the point is that mereotopology by itself says nothing specific about *which* parts of the complement qualify as holes. The boundaries of a hole simply cannot be determined by purely mereological or topological considerations. Of course, we have seen that mereotopology says nothing about the boundaries of material objects either. But draw such boundaries as you like, chose the objects you like, unless you also draw the boundaries of their holes (if any) you cannot get a full picture of the mereotopological structure of the objects themselves. And to draw the boundaries of something is to confer ontological dignity to it.

In Casati and Varzi (1994, 1999) it is argued that this alternative way of construing the limits of mereotopology has far reaching consequences. Suppose we take holes seriously: a hole in an object is something with well-defined boundaries. Then the fine-grained distinctions mentioned above can be recovered

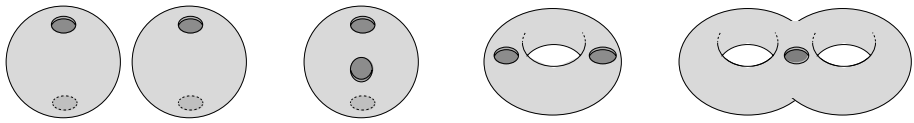


Figure 15.22. The objects of Fig. 15.21 have topologically different internal skins.

by looking at the mereotopological interplay between matter and void, at the properties of the boundary where an object comes into contact with its holes. More precisely, let the interface between two entities x and y be the product of their boundaries:

(114) $x|y =_{df} \mathsf{b}x \times \mathsf{b}y$ Interface

And let the internal skin of an object x be the interface between x and the fusion of its holes. Using 'H' for the binary relation 'is a hole in', this can be defined as follows:

(115) $\mathsf{s}x =_{df} \Sigma z \exists y(\mathsf{H}yx \wedge z = y|x)$ skin

Then the distinctions in question are distinctions that reflect the mereotopology of an object's skin. With reference to Fig. 15.21, for example, it can be checked that the skin of the doubly perforated block on the left is the disconnected sum of two cylinders, i.e., topologically, two spherical surfaces with two punctures each. By contrast, the block with a Y-shaped hole has a connected skin that is equivalent to a spherical surface with three punctures, while the other blocks have skins equivalent to a torus with two punctures and to a bitorus with one puncture, respectively (Fig. 15.22). (Note that a puncture is not a hole but a mere boundary. The surfaces of the objects in Fig. 15.21 do not have boundaries, yet their internal skins do—and that makes all the difference.)

Now, in a boundaryless theory all of this is beyond reach. But in a sufficiently strong boundary-based theory such as $KGEMT$ the notion of an internal skin is perfectly meaningful and well defined for every object of positive genus, and its mereotopological classification does not present any special challenge. This confirms once again the greater expressive power that comes with an ontological commitment to boundaries. It also shows, however, that such a commitment is not enough: the existential quantifier in definition (115) shows that an explicit commitment to holes is also needed. To the extent that the binary predicate 'H' is to be treated as a primitive, it is clear that this requires a step beyond $KGEMT$ and its pure mereotopological extensions. (For an axiomatic treatment of 'H' and of its interplay with 'P' and 'C', see Casati and Varzi 1994, Appendix, and Varzi 1996b.)

3. Location theories

Let us finally turn to the relation of spatial location. Intuitively, this is the relation that holds between an entity and the spatial region that it occupies, and we have already seen that this relation can hardly be reduced to a chapter of mereology and/or topology. Even if it were—as someone inclined to favor a Leibnizian, relationist conception of space against its Newtonian, substantivalist foes would urge—methodological prudence would suggest that we regard the reduction as a theorem, not as a starting point, hence that the relation of location be treated as an independent primitive next to parthood and connection. Exactly how this relation should be characterizeed, and how it should interact with the principles governing those other primitives, is precisely the sort of question that a good theory of location should aim to answer.

3.1 Varieties of Location

Before looking at the main options, the usual terminological caveats are in order. As with 'part of' and 'connected to', locative predicates have various meanings in ordinary language and it is important to be explicit.

For one thing, we often speak so as to specify the location of an object by reference to another object, as opposed to a spatial region. Consider:

(116) The biceps muscle is located in the arm.
(117) The parking area is located next to the stadium.
(118) The elevator is located inside the main building.

Pretty clearly, such cases are of no special interest, as they reflect different ways of asserting mereotopological relations of the familiar sort: in (116) the locative predicate is just a variant of 'part of', in (117) it expresses the relation of external connection, and in (118) it stands for a relation of containment that can be cashed out in terms of interior proper parthood. Of course, establishing mereotopological relations may be an *indirect* way of specifying a genuine location: insofar as the biceps muscle is part of the arm, for instance, the muscle is bound to be located within the region occupied by the arm, though this is by itself an intuition that needs to be spelled out carefully (Sec. 3.3). Moreover, not every case of relative location can be explained in this fashion (Sec. 3.4). For the moment, let us just emphasize that the main concern of a theory of spatial location as we understand it here is with those cases in which an object's location is specified *directly*, as in

(119) The peak of Mount Everest is located at $27°59'$ N $86°56'$ E.
(120) The new library will be located at this site.

It is in this sense that the theory presupposes an ontology that includes spatial regions as *bona fide* entities in their own right. Indeed, we shall assume that

the location primitive is a relation whose second argument can *only* be a region of space—a 'place'. (Never mind the question of what sort of linguistic expressions can serve the purpose of referring to places, as opposed to things that *have* a place. Statements such as

(121) The bookcase is located in the living room.
(122) The United Nations are located in Manhattan.

are somewhat ambiguous in this respect, but we may suppose that the context will always suffice to determine the intended meaning.)

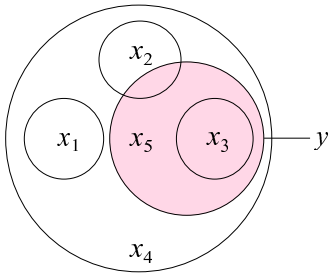
Secondly, there are various ways in which an object may be said to be located at a region. In a very loose sense, I am located at any region that is not completely free of me (this room, or even the adjacent dining room if I am reaching a foot out of the doorway); in a stricter sense, I am only located at those regions that host me entirely (this room, if I am not reaching out of the doorway); and in a stricter sense still, I am only located at one region, namely the region that corresponds exactly to the volume of my body. In the following we shall use 'located at' as designating the last, strictest relation; the weaker relations can be introduced by definition. More precisely, suppose we use 'L' for the predicate of exact location. Then three additional predicates can immediately be defined as follows (from Parsons 2006[14]):

(123) $\mathsf{GL}xy =_{df} \exists z(\mathsf{O}zy \wedge \mathsf{L}xz)$ Generic Location
(124) $\mathsf{EL}xy =_{df} \exists z(\mathsf{P}zy \wedge \mathsf{L}xz)$ Entire Location
(125) $\mathsf{UL}xy =_{df} \exists z(\mathsf{P}yz \wedge \mathsf{L}xz)$ Ubiquitous Location

Thus, I am generically, in fact entirely located in Manhattan, but not ubiquitously (or exactly) located there; I am generically, in fact ubiquitously located at the region occupied by my left arm, but not entirely (or exactly) located there; and if I reach an arm in my neighbor's window, then I am generically, but neither entirely nor ubiquitously (let alone exactly) located at the region corresponding to her living room. As we shall see, under suitable conditions these predicates are interdefinable, so the choice of 'L' as a primitive is ultimately immaterial.

Finally, it goes without saying that the location of an object may change over time. This would suggest treating 'located at', not as a binary predicate, but as a three-place predicate involving a spatial as well as a temporal argument (or as a temporally indexed binary predicate). However, we have seen that the same goes for parthood and connection: unless one accepts a radical form of mereotopological essentialism, an object may in principle change its parts or its topological relations without ceasing to exist. In the preceding sections we

[14]Parsons's term for generic location is 'weak location', and his term for ubiquitous location is 'pervasive location'. Our different terminology is dictated merely by notational convenience, in view of the predicates 'whole location' (WL) and 'proper location' (PL) introduced below.



	$\mathsf{GL}x_ny$	$\mathsf{EL}x_ny$	$\mathsf{UL}x_ny$	$\mathsf{L}x_ny$
$n = 1$	–	–	–	–
$n = 2$	+	–	–	–
$n = 3$	+	+	–	–
$n = 4$	+	–	+	–
$n = 5$	+	+	+	+

Figure 15.23. Basic locative relations.

have tried to keep things simple by treating ‘P’ and ‘C’ as binary predicates, and we shall do the same with our location primitive ‘L’. In a way, this means that we are assuming the relevant time to be fixed throughout. But one may also consider an alternative reading, to the effect that the variables of the theory range over four-dimensional entities extended in space-time. (See again the brief discussion following (32), Sec. 1.3.2.) Not much of what follows depends on the strategy one favors, but for simplicity we shall continue to speak of the location of an object as a 3D region of space rather than—possibly—a 4D region of space-time.

3.2 Basic Principles

With these conventions in place, let us officially expand our formal language by adding the new binary predicate ‘L’, intuitively understood as the relation of exact location holding between an object and a region of space. To make this interpretation explicit, we may begin by assuming the following axiom:

(L.1) $\mathsf{L}xy \wedge \mathsf{L}xz \rightarrow y = z$ Functionality

This guarantees that nothing can have more than one exact location, which is all that is needed to justify the definitions in (123)–(125) (Fig. 15.23.). Indeed, it is easy to see that in the presence of an extensional mereology, (L.1) has the following corollaries:

(126) $\mathsf{L}xy \leftrightarrow (\mathsf{EL}xy \wedge \mathsf{UL}xy)$
(127) $\mathsf{UL}xy \leftrightarrow (\mathsf{GL}xy \wedge \forall z(\mathsf{O}zy \rightarrow \mathsf{GL}xz))$
(128) $\mathsf{EL}xy \leftrightarrow (\mathsf{GL}xy \wedge \forall z(\mathsf{GL}xz \rightarrow \mathsf{O}zy))$

Thus, although we have settled on the strictest possible primitive, the predicate ‘L’, one could equally well settle on the weakest predicate, ‘GL’, and define the rest *via* (126)–(128). More precisely, if the mereological theory in the background is at least as strong as EM, it turns out that the L-based system

defined by (L.1) plus (123)–(125) is equivalent to the GL-based system defined by (126)–(128) plus the following:

(L.2) $\mathsf{GL}xy \rightarrow \exists z \mathsf{L}xz$ Exactness

(See Parsons 2006; for a different choice of primitives, compare also Perzanowski 1993.) Note that the conjunct '$\mathsf{GL}xy$' is redundant in (127) as long as O is reflexive. However, this extra conjunct is needed in (128) unless one assumes that everything is located somewhere:

(L.3) $\exists y \mathsf{L}xy$ Spatiality

This is clearly an assumption that reflects a substantive thesis (a central tenet of most nominalistic ontologies), so it is fair to keep it separate.

To be sure, there is a sense in which (L.1) may also be read as a substantive thesis: functionality is reasonable only to the extent that we are thinking of so-called 'particular' entities, entities such as material bodies or events, as opposed to 'universal' entities such as properties. That a material body cannot be in two places at once was already a central thesis of Aristotle's theory of location (Morison 2002). But Aristotle also held the view that universals, too, exist in space and time: they exist wherever and whenever they are exemplified. Wisdom, for example, exists whenever and whevever there are wise people—and whenever a wise person exists, wisdom exists in its entirety wherever that person is located. Wisdom can therefore be multi-located, and the same goes for all universals. Since this view is still very popular (Armstrong 1989), the functionality principle (L.1) would be objectionable. However, we can bypass this issue by taking 'L' to represent the location relation that is characteristic of particulars. In that sense, (L.1) is, if not a conceptual truth, a perfectly reasonable starting point.

What else is needed in order to fix the intended meaning of 'L' in this sense? Since the idea is that every object must be located at a *region*, some restriction must be imposed on the second argument of the relation. This would be a trivial task if the language contained an explicit predicate for regions. However, we can make do without such a predicate and try to characterize regionhood directly in terms of suitable axioms on L. There are two options here, depending on whether we think that spatial regions are themselves entities located somewhere. If we think so, then the obvious thing to say is that such entities can only be located at themselves (Casati and Varzi 1999, p. 121):

(L.4) $\mathsf{L}xy \rightarrow \mathsf{L}yy$ Conditional Reflexivity

This would immediately imply that no distinct regions can be exactly co-located, i.e., effectively, located at each other:

(129) $\mathsf{L}xy \wedge \mathsf{L}zw \wedge \mathsf{L}yw \rightarrow y = w$

Moreover, given (L.1), conditional reflexivity would ensure that L is both antisymmetric and transitive:

(130) $\mathsf{L}xy \wedge \mathsf{L}yx \rightarrow x = y$
(131) $\mathsf{L}xy \wedge \mathsf{L}yz \rightarrow \mathsf{L}xz$

It follows that relative to the sub-domain of regions L would behave as a partial ordering. By contrast, if we think that regions do not have a location—they *are* locations—then the obvious option is given by:

(L.5) $\mathsf{L}xy \rightarrow \neg\mathsf{L}yz$ Conditional Emptiness

In this case, the restriction of L to the class of regions would again qualify as a partial ordering—a strict ordering—but only in a trivial sense: effectively, it would just collapse to the empty relation.

There is, arguably, no deep metaphysical issue behind these two options: both (L.4) and (L.5) are equally good stipulations, and the difference would disappear as soon as we focus on cases of *proper* location:

(132) $\mathsf{PL}xy =_{df} \mathsf{L}xy \wedge \neg\mathsf{L}yx$ Proper Location

However, there are some differences that are worth mentioning. For one thing, given (L.1), the first option makes it possible to define regionhood in a perfectly straightforward way:

(133) $\mathsf{R}y =_{df} \exists x\mathsf{L}xy$ Region

(We are speaking of regions in a broad sense, including boundaries as limit cases.) By contrast, (L.5) would support this definition—and variants thereof—only on the assumption that there are no unoccupied regions, i.e., regions that fail to correspond to the location of some object or event. To put it differently, if all location is proper location, it is not possible to define regionhood unless the following principle is accepted:

(L.6) $\mathsf{R}y \rightarrow \exists x\mathsf{PL}xy$ Fullness

And philosophically this principle is just as controversial as the spatiality principle (L.3). (Among other reasons, one might want to allow for boundary-like regions while rejecting the existence of boundary-like objects, as seen in Sec. 2.4.3.) Secondly, it is also apparent that the two options differ with regard to (L.3) itself: (L.4) is compatible with this principle, (L.5) isn't. Thus, the second option makes it impossible to assert the thesis that everything is located somewhere—a thesis which, albeit controversial, is certainly not inconsistent. Again, this is a limitation that would dissolve if 'R' were available in the language, in which case the thesis in question could be reformulated as follows:

(L.3′) $\neg \mathsf{R}x \rightarrow \exists y \mathsf{L}xy$

But precisely because ‘R’ cannot be defined absent (L.6), the limitation is not immaterial.

For these reasons, in the following we shall favor the first option and assume the conditional reflexivity axiom (L.4). Together with the functionality postulate (L.1), this yields a minimal theory of (exact) spatial location, which we shall label S: this theory is incompatible with (L.5), but it includes the exactness principle (L.2) as a theorem and can be strengthened by adding the spatiality principle (L.3), the fullness principle (L.6), or both.

At this point, we could consider various ways of strengthening S (or its extensions) by imposing suitable axioms on the predicate ‘R’ defined in (133). For example, it seems reasonable to assume that regionhood is both dissective and cumulative, i.e., that any part of a region is itself a region, and that the sum of any regions (if it exists) is a region, too:

(L.7) $\mathsf{R}y \wedge \mathsf{P}xy \rightarrow \mathsf{R}x$ Dissectiveness

(L.8) $z = \Sigma x \phi x \wedge \forall x(\phi x \rightarrow \mathsf{R}x) \rightarrow \mathsf{R}z$ Cumulativity

It may also be reasonable to consider additional postulates concerning whether the class of all regions forms a dense domain, or whether it forms an atomless, possibly a boundaryless domain as opposed to an atomistic domain every element of which consists (intuitively) of points:

(L.9) $\mathsf{R}x \wedge \mathsf{R}y \wedge \mathsf{PP}xy \rightarrow \exists z(\mathsf{R}z \wedge \mathsf{PP}xz \wedge \mathsf{PP}zy)$ R-Density

(L.10) $\mathsf{R}x \rightarrow \exists y(\mathsf{R}y \wedge \mathsf{PP}yx)$ R-Atomlessness

(L.11) $\mathsf{R}x \rightarrow \exists y(\mathsf{R}y \wedge \mathsf{IPP}yx)$ R-Boundarylessness

(L.12) $\mathsf{R}x \rightarrow \exists y(\mathsf{R}y \wedge \mathsf{A}y \wedge \mathsf{P}yx)$ R-Atomicity

More generally, it may be reasonable at this point to consider whether the domain of regions should be closed under various mereotopological principles, regardless of whether such principles hold of the entities that may occupy those regions. For example, even an anti-extensionalist about material objects will presumably deny that different regions may consist of the same proper parts, and even those who have misgivings about strong composition principles for arbitrary obects might be happy to endorse unrestricted composition of spatial regions. All such extensions of S are obviously worth examining, and they are crucial if we want to fix the intended range of the relational predicate ‘L’, but there is no need here to review all the options: suffice it to say that the availability of ‘R’ makes it possible to examine them in a systematic fashion. (An interesting question, for instance, is whether one can provide a purely mereotopological characterization of Euclidean space; see Tsai 2005, §7.4, for a negative answer). Rather than focusing on the structure of space *per se*, let us see how S can be further extended by considering more closely the relationship

between the two terms of the location relation—the structure of regions *and* the structure of their tenants.

3.3 Mirroring Principles

To this end, let us begin by noting that the four relations L, GL, UL, and EL do not exhaust all the options. Additional locative relations can be specified by replacing the plain mereological predicates in (123)–(125) with finer-grained mereotopological predicates—for example:

(134) $\mathsf{TEL}xy =_{df} \exists z(\mathsf{TPP}zy \wedge \mathsf{L}xz)$ Tangential EL
(135) $\mathsf{IEL}xy =_{df} \exists z(\mathsf{IPP}zy \wedge \mathsf{L}xz)$ Interior EL
(136) $\mathsf{TUL}xy =_{df} \exists z(\mathsf{TPP}yz \wedge \mathsf{L}xz)$ Tangential UL
(137) $\mathsf{IUL}xy =_{df} \exists z(\mathsf{IPP}yz \wedge \mathsf{L}xz)$ Interior UL

More generally, given any mereotopological relation ψ, there is a corresponding locative relation L_ψ defined by:

(138) $\mathsf{L}_\psi xy =_{df} \exists z(\psi zy \wedge \mathsf{L}xz)$ ψ-Location

(Thus, $\mathsf{GL} = \mathsf{L}_\mathsf{O}$, $\mathsf{EL} = \mathsf{L}_\mathsf{P}$, $\mathsf{UL} = \mathsf{L}_{\check{\mathsf{P}}}$, etc.) Such generalizations are straightforward, but they bear out that the language of location can be as rich as the underlying mereotopological vocabulary. Indeed, this is only half of the story. According to (138), to be ψ-located at a region y is to be exactly located at some region, z, that is ψ-related to y. Thus, the resulting variety of locative relations is defined with reference to the mereotopological structure of the *range* of L—the structure of space. But one could also consider the obvious alternative, and characterize locative relations by reference to the mereotopological structure of the *domain* of L—the structure of space's tenants. In this alternative sense, to be ψ-located at a region y is to be ψ-related to some object, z, that is exactly located at y:

(139) $\mathsf{L}^\psi xy =_{df} \exists z(\psi xz \wedge \mathsf{L}zy)$ ψ-Location (2)

Now, the interesting question is whether these two ways of characterizing locative relations should coincide—whether 'L^ψ' and 'L_ψ' should always stand and fall together. This is trivially true when the relata are of the same kind, i.e., regions, for in that case locative relations collapse to mereotopological relations in view of the following S-theorems:

(140) $\mathsf{R}x \rightarrow (\mathsf{L}_\psi xy \leftrightarrow \psi xy)$
(141) $\mathsf{R}x \rightarrow (\mathsf{L}^\psi xy \leftrightarrow \psi xy)$

But what about the general case? This is not just an interesting question to ask if we want to get the map straight; it is also a question that calls for interesting philosophical decisions.

To address the matter properly, let us consider the two directions of the equivalence separately, corresponding to the following theses:

(L.13) $\mathsf{L}^{\psi}xy \rightarrow \mathsf{L}_{\psi}xy$ Bottom Mirroring
(L.14) $\mathsf{L}_{\psi}xy \rightarrow \mathsf{L}^{\psi}xy$ Top Mirroring

3.3.1 Bottom mirroring and co-location. Informally, the first of these theses says that the structure of space should be at least as rich as the structure of its tenants For example, suppose I am exactly located at region r. Then my right foot, which is part of me, is L^{P}-located at r. But if my foot is part of me, then it is reasonable to suppose that its exact location is part of my exact location, which is precisely what (L.13) would imply: my foot is L_{P}-located (i.e., entirely located) at r. For another example, since my foot is connected to the rest of my body, which is located at a certain region r', then it is reasonable to suppose that the location of my foot is connected to r', too: my foot being L^{C}-located at r' implies its being L_{C}-located at r'. In general, adding (L.13) to S would secure that if two objects are ψ-related (where ψ is P, C, or any other mereotopological relation), then so are their locations:

(142) $\mathsf{L}xy \wedge \mathsf{L}zw \wedge \psi xz \rightarrow \psi yw.$

Plausible as all this might sound, it is however easy to see that (L.13) does not generally hold. A simple counterexample is depicted in Fig. 15.24, left. In this model there is just one (atomic) region, r, and a complex object, a, consisting of two (atomic) parts, b and c. Object a is located at r, hence its parts b and c are L^{P}-located at r. Yet they are not L_{P}-located at r because there is no part of r at which they are exactly located. Note that the mereotopological relations in this model can easily be extended so as to satisfy all the axioms of $KGEMT$, so the counterexample does not depend on the strength of the underlying mereotopological theory. It depends esclusively on the behavior of the relation L.

Now, it might be observed that this model (or its $KGEMT$ closure) would be ruled out if we assumed that every part of a spatially located entity had a spatial location—a restricted form of the spatiality principle (L.3). More generally, it might be thought that in order to justify (L.13) one must assume the following:

(L.15) $\psi xz \wedge \mathsf{L}zy \rightarrow \exists w \mathsf{L}xw$ Conditional Spatiality

Bottom mirroring should not hold *holus bolus*, but only when x and its ψ-relata are genuine spatial entities (in which case (L.13) turns out to be eqivalent to (142)). However, it is easy to see that even S + (L.15) would fail to warrant this claim. A counterexample is depicted in Fig. 15.24, center. Here b and c are L^{PP}-located at r, since each is a proper part of a, which is located at r. Yet

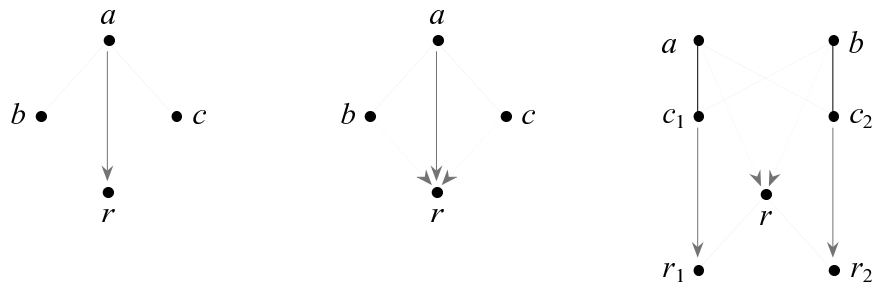


Figure 15.24. Three violations of Bottom Mirroring. (Location relationships are represented by gray arrows; reflexivities omitted.)

there is no proper part of r at which a or b is located, which is to say that neither is L_{PP}-located at r.

We begin here to see the hidden force of (L.13). By requiring that the structure of space mirror the structure of its tenants, this principle rules out the possibility that mereotopologically distinct entities be co-located, i.e., located at the same region. Of course, it may be difficult to imagine a concrete scenario corresponding to the model in question, and certainly very difficult to provide a less abstract representation of it. But it is not difficult to provide a good example if, for instance, we give up mereological extensionality. Consider again Tibbles, the cat, and the 'mere' mereological sum of his tail with the rest of his body, Tib + Tail (Sec. 1.3.2). Obviously, even if one treated these entities as distinct, one would still like to say that they share one and the same location. Yet, if 'ψ' expresses the relevant mereotopological relationship of distinctness *cum* sameness of proper parts, it is obvious that the region to which Tibbles (or Tib + Tail) bears the relation L^{ψ} is not the region to which it bears the relation L_{ψ} (Fig. 15.24, right). Moreover, even in the presence of mereological extensionality one may conceive of situations where spatial co-location seems possible. We have already seen, for instance, that according to Chisholm's (1984) Brentanian theory, topological connection is explained precisely in terms of spatial coincidence of boundaries: boundaries are *located in* space but do not *occupy* space, hence they can coincide while being distinct (Sec. 2.4.2). For a scenario compatible with the full strength of $KGEMT$, consider Davidson's (1969) example concerning event identity: arguably the rotation and the getting warm of a metal ball that is spinning fast are two events, yet they occur exactly in the same region and they share that location with the ball itself. Finally, if our ontology is rich enough to include immaterial or otherwise ethereal creatures for which genuine interpenetration is possible, then again co-location seems conceivable. Already Leibniz mentioned shadows as a case in point (*New Essays*, II-xxvii-1); other candidates include clouds (Shorter 1977), holes

(Casati and Varzi 1994), ghosts (van Inwagen 1990, p. 81), and even angels (Lewis 1991, p. 75).

These are just some examples. But they are indicative of the many philosophical motives that may lie behind a rejection of the principle according to which proper spatial co-location is impossible—a principle that can be put thus:

(L.16) $\mathsf{PL}xy \wedge \mathsf{PL}zy \rightarrow x = z$ Exclusiveness

Giving up this principle involves giving up (L.13), at least in its general form. And the question of what special instances one should posit, i.e., what values of 'ψ' and 'x' satisfy bottom mirroring, calls for a detailed case-by-case investigation—and for explicit ontological decisions.

3.3.2 Top Mirroring. Consider now the converse of (L.13), namely the principle of top mirroring, (L.14). Informally, this says that the structure of space should be mirrored in the structure of those entities that inhabit it. For example, if the location of my body properly includes a region r, then it is reasonable to suppose that my body properly includes something located at r: L_{PE}-location, hence L^{PE}-location. Pretty clearly, however, there are numerous relations ψ for which this sort of implication appears problematic. The location of my body is a proper part of any region that includes this room, but there is no obvious reason to think that every such region is the location of some existing object—no reason to think that L_{PP}-location implies L^{PP}-location. (Compare Fig. 15.25, left.) Similarly, my body is L_{EC}-located at many regions, viz. regions externally connected to my body's current location; yet there is no obvious reason to think that my body is L^{EC}-located at those regions, since many of them may be (partly) empty (Fig. 15.25, center.) Just as bottom mirroring appears to presuppose the spatiality principle, or at least its restricted variant (L.15), top mirroring appears to presuppose fullness, or at least the following restricted version:

(L.17) $\mathsf{R}y \wedge \psi zy \wedge \mathsf{L}xz \rightarrow \exists w \mathsf{PL}wy$ Conditional Fullness

And this is a substantive presupposition that few might grant, regardless of their views concerning purely mereotopological matters.

Indeed, even when ψ is the seemingly innocent relation of proper extension, as in our first example, (L.17) appears problematic. Misgivings about this principle—hence about the corresponding instance of top mirroring—come in various forms. First of all, there are arguments that purport to show that the principle is empirically false (Parsons 2000). Second, the principle rules out *a priori* the possibility of spatially extended mereological atoms, as in Fig. 15.25, right. To the extent that one can conceive of such things, it is argued, it should not be a conceptual truth that every region ubiquitously occupied by an object is exactly occupied by a part of that object. (See Markosian 1998a; other

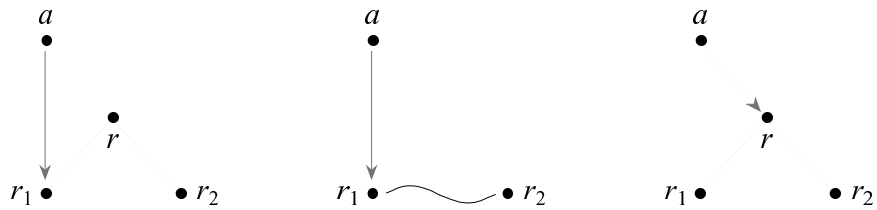


Figure 15.25. Three violations of Top Mirroring.

contemporary philosophers who endorse the possibility of extended simples include Parsons 2004 and Simons 2004, but the view goes back to Democritus's claim that atoms come in an infinite variety of shapes and sizes.) Third, there are arguments to the effect that (L.17) sits ill with the thought that ordinary material bodies can gain or lose some parts (van Inwagen 1981). To illustrate, consider again Tibbles, the cat whose tail gets annihilated at t, and suppose we agree that it survives the accident. Prior to t, (L.17) would suggest that in addition to the whole cat there exist also two externally connected proper parts: Tail and Tib (the remainder). Now consider the following statements:

(143) Tib (before t) = Tib (after t)
(144) Tib (after t) = Tibbles without Tail (after t)
(145) Tibbles without Tail (after t) = Tibbles with Tail (before t)
(146) Tibbles with Tail (before t) $\neq$ Tib (before t)

These four statements are jointly inconsistent, unless one is willing to give up the transitivity of identity, so something must give. (146), however, is trivially true: there is no way one could identify a whole cat with its tailless portion. And (145) is true by assumption: to give it up is to deny that Tibbles survives the accident, unless one is willing to construe cats as four-dimensional entities whose temporal parts are numerically distinct (Heller 1984). As for (144), its denial would obviously incur in a commitment to properly co-located entities, let alone a violation to mereological extensionality (Wiggins 1968). Thus—the argument goes—unless one is ready to accept such unpalatable consequences, the only option is to give up (143): that identity is false for the simple reason that prior to the accident Tib does not even exists; it only exists after the accident, and it exists as tailless Tibbles.[15] (One might also say that before the accident Tib does not *actually* exist. The view that undetached parts are mere 'potential entities' has been the focus of an intense debate in early modern philosophy; see

[15]The puzzle raised by (143)–(146) has been introduced to contemporary discussion by Wiggins (1968), but it goes back at least to the Stoics; see e.g. Sorabji (1988, §1.6). For a detailed overview, see Simons (1987, §3.3) and the introduction to Rea (1997).

Holden 2004. Brentano 1933 endorsed a similar view, too, and some authors have applied it explicitly to the puzzle in question—e.g., Smith 1994, §3.5.)

Whether any such arguments are found compelling is, of course, an open issue. Nonetheless, the obvious moral is that (L.14) can hardly be regarded as a conceptual truth about location. Even if we confine ourselves to its single, *prima facie* plausible instance in which ψ is the relation of (proper) extension, i.e., the converse of (proper) parthood, top mirroring is a substantive metaphysical thesis whose addition to S must be independently motivated.

3.3.3 Further locative relations. In discussing such matters, it is useful to keep in mind that the lack of a full correspondence between the mereotopology of space and the mereotopology of its tenants may find expression in the failure of other principles or equivalences that might otherwise suggest themselves. Consider, for instance, the following relation (from Parsons 2006, §4):

(147) $\mathsf{WL}xy =_{df} \forall z(\mathsf{P}zx \rightarrow \mathsf{GL}zy)$ Whole Location

Intuitively, this says that an object is located in this room (for instance) if every part of the object is generically located in this room, i.e., if none of it is missing from the room. This might sound like a different way of saying that the object is entirely located in this room, in the original sense of (124) (corresponding to L_{P}-location) or in the alternative sense of (139) (L^{P}-location). In fact, however, all these notions are distinct. Not only do L_{P}-location and L^{P}-location come apart, as seen above. They also differ from whole location: the diagram in Fig. 15.24, left, corresponds to a model in which an object, a, is both L_{P}- and L^{P}-located at a region, r, in spite of not being wholly located there; the diagram in Fig. 15.25, right, depicts a model in which a (an extended atom) is wolly located at region r_1 in spite of being neither L_{P}- nor L^{P}-located there.[16] (Note that both of these models are extensional, and would continue to exhibit these features even if closed under every $KGEMT$-axiom.)

The notion of whole location is just one example. In general, for any mereotopological relation ψ, one can define two additional locative relations besides L_{ψ} and L^{ψ}, obtained by switching to a universal quantifier and replacing 'L' by 'GL':

(148) $\mathsf{L}_{\forall\psi}xy =_{df} \forall z(\psi zy \rightarrow \mathsf{GL}xz)$

(149) $\mathsf{L}^{\forall\psi}xy =_{df} \forall z(\psi xz \rightarrow \mathsf{GL}zy)$

Here $\mathsf{WL} = \mathsf{L}_{\forall\breve{\mathsf{P}}}$, and it should be obvious from our single example that the equivalence between the new predicates in (148)–(149) and the old predicates

[16] In Casati and Varzi (1999, §7.2), entire location is called 'whole location' and labelled 'WL'.

in (138)–(139) is generally an open question. Indeed, at this point we get the full picture by further generalizing these four basic patterns in the obvious (recursive) way: if λ is *any* locative relation, then so are the following:

(150) $\mathsf{L}_{\exists\psi\lambda}xy =_{df} \exists z(\psi zy \wedge \lambda xz)$
(151) $\mathsf{L}^{\exists\psi\lambda}xy =_{df} \exists z(\psi xz \wedge \lambda zy)$
(152) $\mathsf{L}_{\forall\psi\lambda}xy =_{df} \forall z(\psi zy \rightarrow \lambda xz)$
(153) $\mathsf{L}^{\forall\psi\lambda}xy =_{df} \forall z(\psi xz \rightarrow \lambda zy)$

There are lots of redundancies and empty relations in this picture, whose complexity depends significantly on the mereotopological axioms governing ψ. Nonetheless, it is only through a careful study of such intricacies—and of the corresponding mirroring principles, at the moment vastly unexplored—that a reasonably complete theory of location can emerge.

3.4 Relative locations

The axiomatization of the region predicate 'R' and the positing of suitable mirroring principles constitute the two main directions in which theory S can be extended. Let us briefly mention a third direction, whose ramifications span philosophical and methodological issues alike. We have said that in ordinary language location is often understood as a relation between two objects, as opposed to an object and a spatial region, and we have said that such understandings need not be taken to express any fundamental relationships: often, such 'relative' locations are mereotopological relations in disguise, as in examples (118)–(120). There are, however, cases that resist this sort of explanation. Consider:

(154) The brain is located inside the cranial cavity.
(155) The swimming pool is located behind the house.
(156) The bus stop is located right across the old oak tree.

Surely the truth conditions of these statements can hardly be explained in terms of mereotopological relations, so one can hardly leave it at that. Can S be strengthened so as to account for such cases as well?

In a way, the answer is straightforward. Consider (154). Although there is no direct mereotopological relationship between the brain and the cranial cavity (unless one thinks the latter literally surrounds the former), one can still explain the relevant truth conditions by reference to the mereotopology of the corresponding spatial locations: statement (154) is true if, and only if, the location of the brain is an interior proper part of the location of the cranial cavity. Equivalently, (154) is true if and only if the brain is located entirely in the interior of the spatial region occupied by the cranial cavity. This suggests that cases such as this can easily be accommodated within the present framework.

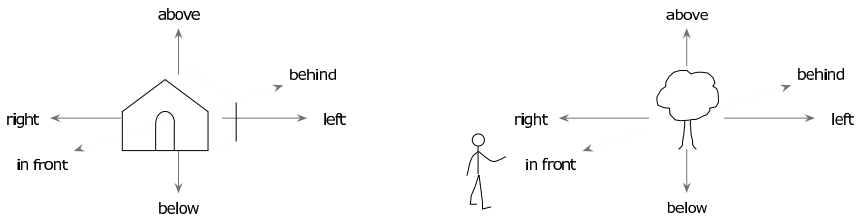


Figure 15.26. Object-centered (left) and observer-centered (right) frames of reference.

When we say that a 'target' object, x, bears a certain locative relation to a 'reference' object, y, we mean to say that x bears that relation to y's place. To make this clear, let us introduce an explicit location functor, whose uniqueness follows directly from the functionality axiom (L.1):

(157) $\mathsf{p}x =_{df} \imath y \mathsf{L} xy$ place

(See Donnelly 2004 for a location theory with p treated as a primitive.) Then, for any locative relation λ and any spatial objects x and y, we can define a corresponding predicate of relative location '$\mathsf{R}\lambda$' as follows:

(158) $\mathsf{R}\lambda xy =_{df} \lambda x(\mathsf{p}y)$ Relative Location

In the brain-cavity case, λ is the relation IEL defined in (135), i.e., $\mathsf{L}_{\mathsf{IPP}}$, but the same pattern would apply to a large variety of other cases, provided the reference object has a location somewhere. In fact, the same pattern can be applied to account for the initial examples in (116)–(118), too: the biceps muscle is RL_{P}-located at the arm; the parking area is $\mathsf{RL}_{\mathsf{EC}}$-located at the stadium; the elevator is $\mathsf{RL}_{\mathsf{IPP}}$-located at the main building. In the presence of suitable bottom mirroring conditions, these three claims are equivalent to the pure mereotopological claims obtained by replacing the relation RL_{ψ} with ψ itself.

Cases such as (155) and (156) are different. Here the difficulty does not just lie in the fact that the target object and the reference object do not stand in any mereotopological relation to each other. The difficulty is that the spatial relationships reported by such statements—corresponding to such prepositions as 'behind' and 'across', but also 'above', 'underneath', 'left of', etc.—have little or nothing to do with mereotopology. There is in fact a large literature devoted to the semantics of spatial prepositions (beginning with the classic work of Herskovitz 1986) and it is fair to say that their treatment requires a degree of sophistication that goes far beyond the conceptual apparatus developed above. Among other things, there are well-known complications owing to the fact that their treatment calls for a systematic distinction between object-centered frames of reference, as in (155), and observer-centered frames, as in (156) (see

Fig. 15.26), whereas mereotopological relations are completely independent of subjective or perspectival considerations. Still, this is not a limitation that speaks against the employment of a primitive such as 'L'. It is, rather, an indication that the ensuing theory, S, may have to be matched with a more sophisticated background than a mereotopological theory can afford. Thus, suppose we allow the relational predicate 'ψ' in the definitions of Sec. 3.3 to stand for relations that are not purely mereotopological: relations such as 'behind', 'across', etc. (in each of their multiple uses). Then the idea illustrated with reference to the brain-cavity example can in principle be applied also to (155), (156), and the like. To say that the swimming pool is located behind the house is to say that it is $\mathsf{RL}_{\mathsf{behind}}$-located at the house. To say that the bus stop is located across the tree is to say that it is $\mathsf{RL}_{\mathsf{across}}$-located at the tree. And so on. Exactly how these relations should be formalized, i.e., what principles should be posited to fix the logical behavior of the relevant ψ, is where things get difficult and may require a lot of detailed work. (We have, after all, seen how difficult it is to do this when ψ is a mereotopological relation.) But that is not to say that relative locations require a *sui generis* treatment that a suitable extension of S could not accommodate.

On the other hand, here is where the main metaphysical assumption underlying S may be questioned. The accommodation comes with definition (158), which allows one to handle relative locations in terms of a primitive relation, L, whose range consists exclusively of spatial regions. In other words, it allows one to express a spatial relation between a target and a reference object as a relation between the target and the reference's *place*. But one might object that this has things the wrong way round. Relative locations are ontologically neutral with respect to the status of space, whereas the proposed treatment depends crucially on the assumption that places, and spatial regions generally, are entities of a kind. A relationist about space might therefore reject the account and take the opportunity to reverse the order of the explanation: to be RL_{ψ}-related to a place is to be ψ-related to the place's tenant, and talk about places is shorthand for talk about spatial objects. Also a substantivalist about space might think that when it comes to spatial reasoning, objects are conceptually prior to their locations, since we cannot identify the latter independently of the former (Strawson 1959, ch. 1.) Even a forerunning substantivalist such as Newton emphasized that absolute places, defined as in (157), are scarcely useful for locating things in the world: we do not locate an object on a moving ship with reference to an immobile environment but, rather, with reference to the ship itself (*Principia*, Definitions, Sch. 4). In short, there may be philosophical as well as methodological reasons for resisting the treatment of locative relations indicated above, and if these reasons are taken seriously, then cases such as (154)–(156) run afoul of the basic framework outlined here and call for independent treatment. An articulated proposal in this spirit may be found in

Donnelly (2005), but much recent literature devoted to the formal representation of direction and other qualitative spatial relations (from the works of Mukerjee and Joe 1990, Frank 1992, and Hernández 1994 to more interdisciplinary works such as those collected in van der Zee and Slack 2003) may be viewed in this perspective.

3.5 Location, connection, and parthood

Let us conclude with a few remarks about the whole conceptual package that we have been putting together. We have at this point three main primitive notions: location, connection, and parthood. Some structural relationships among these notions have been examined, but the general question of their ontological intertwining is still open. Generally speaking, parthood and connection are independent from each other, unless one accepts the converse monotonicity principle (C.4) (Sec. 2.3.2). But what about location? Let us keep with the assumption that location is a relation between a thing and its place. Is that relation completely autonomous or does it entail a mereotopological linkage of some sort?

A purely mereological linkage seems out of the question. There is no reason to think that I share any parts with the space I occupy, just as there is no reason to think that movement—change of location —is a form of mereological change. This is not to say that location implies disjointness, since the first argument of L may itself be a region, or a hybrid fusion including regions among other things. In general, however, it seems perfectly reasonable to assume that the implication holds for *ordinary* cases of location—a thesis that can be put as follows:

(L.18) $\mathsf{PL}xy \rightarrow \exists z(\mathsf{P}zx \wedge \mathsf{L}zy \wedge \mathsf{D}zy)$ Spatial Disjointness

What about a topological linkage? In this case the picture seems different. If an object is located at some place, then it might be plausible to suppose that the object and its place are connected in some way:

(L.19) $\mathsf{L}xy \rightarrow \mathsf{C}xy$ Spatial Connection

In what way would I be connected to my place? Since overlap is excluded, the relevant linkage would have to be one of external connection. However, this means that the plausibility of (L.19) depends crucially on the interpretation of the topological primitive 'C'.

If we go along with the standard interpretation, corresponding to a boundary-based theory such as $KGEMT$, the prospects are slim. On that interpretation, two things can be externally connected only if one is open and the other closed, at least in the relevant contact area: two closed entities, or two open entities, can only connect through mereological overlap (Sec. 2.4.1). Now, suppose I am a

closed body. Then (L.19), together with (L.18), imply that the region at which I am located—my place—must be open. That region, however, has a closure, and one should think that the closure of a spatial region is itself a region: its boundary is pure space, too. But my boundary is not pure space: whatever it is, it is part of me, and none of me is made of space. Thus, the closure of my place and I are disjoint—which is to say that we are not connected after all. A different way of putting this involves the thought that my place and I must have the same topology: if I am closed (for instance), then my place must be closed too, hence we cannot be externally connected. Strictly speaking, this involves an appeal to the principle of bottom mirroring, specifically the following instance of corollary (142):

(159) $\mathsf{L}xy \wedge \mathsf{L}zw \wedge \mathsf{IPP}xz \rightarrow \mathsf{IPP}yw$

My interior (x) is a proper interior part of me (z), hence its place (y) must be an interior proper part of my place (w). We have seen that bottom mirroring may not generally hold, but this specific instance seems fair. Yet (L.19) would require the opposite: it would require my place to be an interior proper part of the place of my interior—absurd.

By contrast, suppose we go along with a different interpretation. For instance, consider those interpretations that explain connection explicitly in terms of spatial location. We have seen three such interpretations, corresponding to the following necessary and sufficient conditions for two entities x and y to be connected: (i) the place of x is connected to the place of y (Sec. 2.3.2, thesis (74)); (ii) the place of x overlaps the place of y (Sec. 2.4.2, thesis (89)); or (iii) the closure of the place of x overlaps the closure of the place of y (Sec. 2.4.3, thesis (90)). With the help of 'L', and assuming the spatiality axiom (L.3), these three interpretations can now be formally stated as follows:

(160) $\mathsf{C}xy \leftrightarrow \mathsf{C}(\mathsf{p}x)(\mathsf{p}y)$
(161) $\mathsf{C}xy \leftrightarrow \mathsf{O}(\mathsf{p}x)(\mathsf{p}y)$
(162) $\mathsf{C}xy \leftrightarrow \mathsf{O}(\mathsf{cp}x)(\mathsf{cp}y)$

And, plainly, each of these statements is perfectly compatible with the spatial connection principle (L.19). In fact, since C and O are both reflexive, each statement *entails* (L.19) in view of the following S-theorem:

(163) $\mathsf{L}xy \rightarrow \mathsf{p}x = y = \mathsf{p}y$

This is not surprising. After all, these theories establish an intimate relationship between topology and spatiality, and the claim that location implies connection is but one way of making that explicit. One might even go as far as to say that on these theories location *is* connection of a kind: it is the relation of connection that always holds in the special case where one of the relata equals the place

of the other. On the other hand, it is fair to note that all of this depends on the assumption that L is conditionally reflexive (L.4). Should one decide to go for the alternative option (L.5), treating L as a relation whose domain does not contain any spatial regions, (163) would not hold and (L.19) would not follow from (160)–(162). Indeed, on that way of interpreting 'L', the picture would be perfectly reversed: nothing would be connected to its place because places would lack a place of their own, hence (L.19) would be just as unacceptable in these theories as it is in $KGEMT$.

These are just some examples, but they suffice to show that the ontological intertwining between locative and mereotopological concepts cannot be assessed without a clear stand on some very basic semantic issues concerning those concepts. Of course, this is also true of the interplay between mereological and topological concepts, but in the present case the stakes are higher. Some of the options leave room for no intertwining whatsoever; other options trivialize the intertwining by explaining one sort of concept directly in terms of the other. There is, to be sure, some room for compromise. After all, $KGEMT$ is a pretty strong theory and its weaker children are compatible with (L.19). Similarly, there is strictly speaking no constraint to supplement its main alternatives with the explicit equivalences in (160)–(162), so the status of (L.19) is in principle open for those theories, too. One final remark, however, is in order. For suppose we do accept (L.19) together with (L.8), that is, suppose we do establish an overt mereotopological linkage of external connection between a spatial object and its spatial location. Then it follows that no spatial entity will ever qualify as a interior proper part of anything, except for empty regions of space:

(164) $\mathsf{PL}xy \rightarrow \neg\mathsf{IPP}xz \wedge \neg\mathsf{IPP}yz$

This is an immediate consequence of the definition of 'IPP' (Sec. 2.3.1): my 'interior' proper parts, for example, would immediately turn into tangential proper parts by virtue of being connected to something with which I have no parts in common—their places; and the 'interior' proper parts of my place, at least those proper parts that are not empty, would immediately turn into tangential proper parts by virtue of being connected to something with which they have nothing in common—their material guests. This is bad news, for it means that our mereotopology would collapse altogether. Or rather, it means that it would have to be largely re-written by replacing throughout our topological primitive 'C' with the following impure connection predicate:

(165) $\mathsf{RC}xy =_{df} \mathsf{C}xy \wedge (\mathsf{R}x \leftrightarrow \mathsf{R}y)$ Restricted Connection

It is, of course, this notion of connection that we had in mind in setting up $KGEMT$ and its variants. RC never cuts across levels: in order for two entities to be RC-related, both of them or neither of them must be regions of

space. But then we have come around the circle, for the RC-variant of (L.19) is clearly false.

We face, here, the fundamental limit of topology—and mereotopology—as a general theory of space. One way or the other, the *analysis situs* cannot do proper justice to the fact that objects are *situated*, which is why the theory of location is independently needed. One way or the other, spatial reasoning must come to terms with the fundamental metaphysical mystery on which it depends—embedding in space.

Acknowledgments

Parts of this chapter draw on previous material. In particular, Sec. 1 has some overlap with Varzi (2003) while Sec. 2 has some overlap with Chapters 1, 4 and 5 of Casati and Varzi (1999).

References

Aczel, P. (1988). *Non-Well-Founded Sets*. Stanford: CSLI Publications.

Aiello, M. (2000). Topo-distance: Measuring the Difference between Spatial Patterns. In *Proc. JELIA 00*, pages 73–86. Berlin: Springer.

Armstrong, D. M. (1989). *Universals*. Boulder CO: Westview Press.

Asher, N. and Vieu, L. (1995). Toward a Geometry of Common Sense: A Semantics and a Complete Axiomatization of Mereotopology. In *Proc. IJCAI 95*, pages 846–852. San Mateo: Morgan Kaufmann.

Baker, L. R. (1997). Why Constitution Is Not Identity. *J. of Philosophy*, 94: 599–621.

Baumgartner, W. and Simons, P. M. (1994). Brentano's Mereology. *Axiomathes*, 5:55–76.

Baxter, D. (1988). Identity in the Loose and Popular Sense. *Mind*, 97:575–582.

Bennett, B. (1996a). Carving Up Space: Steps Towards Construction of an Absolutely Complete Theory of Spatial Regions. In *Proc. JELIA 96*, pages 337–353. Berlin: Springer.

Bennett, B (1996b). Modal Logics for Qualitative Spatial Reasoning. *Bulletin of the Interest Group in Pure and Applied Logic*, 4:23–45.

Biacino, L. and Gerla, G. (1991). Connection Structures. *Notre Dame J. of Formal Logic*, 32:242–247.

Bigelow, J. (1996). God and the New Math. *Philosophical Studies*, 84:127–154.

Bolzano, B. (1851). *Paradoxien des Unendlichen*. Leipzig: Reclam.

Borgo, S., Guarino, N., and Masolo, C. (1996). A Pointless Theory of Space Based on Strong Connection and Congruence. In *Proc. KR 96*, pages 220–229. San Francisco: Morgan Kaufmann.

Bostock, D. (1979). *Logic and Arithmetic*, volume 2. Oxford: Clarendon.

Brentano, F. (1906), Nativistische, empiristische und anoetistische Theorie unserer Raumvorstellung. Published posthumously in Brentano (1976), pages 164–77.

Brentano, F. (1933). *Kategorienlehre*. Hamburg: Meiner. Ed. A. Kastil.

Brentano, F. (1976). *Philosophische Untersuchungen zu Raum, Zeit und Kontinuum*. Hamburg: Meiner. Ed. by S. Korner and R .M. Chisholm.

Bunge, M. (1966). On Null Individuals. *J. of Philosophy*, 63:776–778.

Bunt, H.C. (1985). *Mass Terms and Model-Theoretic Semantics*. Cambridge: Cambridge U. Press.

Burkhardt, H. and Dufour, C. A. (1991). Part/Whole I: History. In Burkhardt, H. and Smith, B., editors, *Handbook of Metaphysics and Ontology*, pages 663–673. Munich: Philosophia.

Cartwright, R. (1975). Scattered Objects. In Lehrer, K., editor, *Analysis and Metaphysics*, pages 153–171. Dordrecht: Reidel.

Casati, R. and Varzi, A. C. (1994). *Holes and Other Superficialities*. Cambridge MA: MIT Press.

Casati, R. and Varzi, A. C. (1999). *Parts and Places*. Cambridge MA: MIT Press.

Chisholm, R.M. (1978). Brentano's Conception of Substance and Accident. *Grazer Philosophische Studien*, 5:197–210.

Chisholm, R.M. (1984). Boundaries as Dependent Particulars. *Grazer Philosophische Studien*, 10:87–95.

Chisholm, R.M. (1987). Scattered Objects. In Thomson, J.J., editor, *On Being and Saying: Essays for Richard Cartwright*, pages 167–173. Cambridge MA: MIT Press.

Chisholm, R.M. (1993). Spatial Continuity and the Theory of Part and Whole. A Brentano Study. *Brentano Studien*, 4:11–23.

Clarke, B.L. (1981). A Calculus of Individuals Based on "Connection". *Notre Dame J. of Formal Logic*, 22:204–218.

Clarke, B.L. (1985). Individuals and Points. *Notre Dame J. of Formal Logic*, 26:61–75.

Cohn, A. G., Bennett, B., Gooday, J. M., and Gotts, N. (1997). Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica*, 1:1–44.

Cohn, A. G. and Varzi, A. C. (2003). Mereotopological Connection. *J. of Philosophical Logic*, 32:357–390.

Copeland, B.J. (1995). On Vague Objects, Fuzzy Logic and Fractal Boundaries. *Southern J. of Philosophy*, 33 (Suppl.):83–96.

Correia, F. (2005). *Existential Dependence and Cognate Notions*. Munich: Philosophia.

Cruse, D. A. (1979). On the Transitivity of the Part-Whole Relation. *J. of Linguistics*, 15:29–38.

Davidson, D. (1969). The Individuation of Events. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*, pages 216–234. Dordrecht: Reidel.

De Laguna, T. (1922). Point, Line, and Surface as Sets of Solids. *J. of Philosophy*, 19:449–461.

Doepke, F.C. (1982). Spatially Coinciding Objects. *Ratio*, 24:45–60.

Donnelly, M. (2004). A Formal Theory for Reasoning about Parthood, Connection, and Location. *Artificial Intelligence*, 160:145–172.

Donnelly, M. (2005). Relative Places. *Applied Ontology*, 1:55–75.

Donnelly, M. and Smith, B. (2003). Layers: A New Approach to Locating Objects in Space. In *Proc. COSIT 03*, pages 46–60. Berlin: Springer.

Düntsch, I. and Wang, H. and McCloskey, S. (2001). A Relation-Algebraic Approach to the Region Connection Calculus. *Theoretical Computer Science*, 255:63–83.

Eberle, R. A. (1968). Yoes on Non-Atomic Systems. *Noûs*, 2:399–403.

Eberle, R. A. (1970). *Nominalistic Systems*. Dordrecht: Reidel.

Engel, R. and Yoes, M.G. (1996). Exponentiating Entities by Necessity. *Australasian J. of Philosophy*, 74:293–304.

Fine, K. (1975). Vagueness, Truth and Logic. *Synthese*, 30:265–300.

Fine, K. (1982). Acts, Events, and Things. In *Proc. 6th Wittgenstein Symposium*, pages 97–105. Vienna: HPT.

Forrest, P. (1996). From Ontology to Topology in the Theory of Regions. *The Monist*, 79:34–50.

Frank, A. U. (1992). Qualitative Spatial Reasoning about Distances and Directions in Geographic Space. *J. of Visual Languages and Computing*, 3: 347–371.

Gallois, A. (1998). *Occasions of Identity*. Oxford: Clarendon.

Galton, A. (1996). Taking Dimension Seriously in Qualitative Spatial Reasoning. In *Proc. ECAI 96*, pages 501–505. Chichester: Wiley.

Galton, A. (1999). The Mereotopology of Discrete Space. In *Proc. COSIT 99*, pages 251–256. Berlin: Springer.

Galton, A. (2000). *Qualitative Spatial Change*. Oxford: Oxford U. Press.

Galton, A. (2003). On the Ontological Status of Geographical Boundaries. In Duckham, M. *et al.*, editors, *Foundations of Geographic Information Science*, pages 151–171. London: Taylor and Francis.

Galton, A. (2004). Multidimensional Mereotopology. In *Proc. KR 04*, pages 45–54. Menlo Park: AAAI Press.

Geach, P. T. (1949). On Rigour in Semantics. *Mind*, 58:512–522.

Gerstl, P. and Pribbenow, S. (1995). Midwinters, End Games, and Bodyparts: A Classification of Part-Whole Relations. *International J. of Human-Computer Studies*, 43:865–889.

Giritli, M. (2003). Who Can Connect in RCC? In *Proc. KI 03*, pages 565–579. Berlin: Springer.

Goodman, N. (1951). *The Structure of Appearance*. Cambridge MA: Harvard U. Press. Republish: Dordrecht: Reidel, 1977.

Goodman, N. (1956). A World of Individuals. In Bochenski, J. *et al.*, editors, *The Problem of Universals*, pages 13–31. Notre Dame: Notre Dame U. Press.

Goodman, N. (1958). On Relations that Generate. *Philosophical Studies*, 9: 65–66.

Gotts, N. (1994a). How Far CanWe 'C'? Defining a 'Doughnut' Using Connection Alone. In *Proc. KR 94*, pages 246–257. San Mateo: Morgan Kaufmann.

Gotts, N. (1994b). Defining a 'Doughnut' Made Difficult. In C. Eschenbach *et al.*, editors, *Topological Foundations of Cognitive Science*, pages 105–29. U. of Hamburg: Cognitive Science Program, Report No. 37.

Gotts, N. (1996). An Axiomatic Approach to Topology for Spatial Information Systems. Technical report, U. of Leeds: School of Computer Studies.

Gotts, N., Gooday, J. M., and Cohn, A. G. (1996). A Connection-Based Approach to Commonsense Topological Description and Reasoning. *The Monist*, 79: 51–75.

Hudson, H. (2001). *A Materialist Metaphysics of the Human Person*. Ithaca NY: Cornell U. Press.

Heller, M. (1984). Temporal Parts of Four Dimensional Objects. *Philosophical Studies*, 46:323–334.

Hempel, C. G. (1953). Reflections on Nelson Goodman's "The Structure of Appearance". *Philosophical Review*, 62:108–116.

Henry, D. (1991). *Medieval Mereology*. Amsterdam: Grüner.

Hernández, D. (1994). *Qualitative Representation of Spatial Knowledge*. Berlin: Springer.

Herskovits, A. (1986). *Language and Spatial Cognition*. Cambridge: Cambridge U. Press.

Holden, T. (2004). *The Architecture of Matter*. Oxford: Clarendon.

Hudson, H. (2002). The Liberal View of Receptacles. *Australasian J. of Philosophy*, 80:432–439.

Hudson, H. (2004). Simples. *The Monist*, 87:303–451. Edited volume.

Husserl, E. (1900/1901). *Logische Untersuchungen. Zweiter Band*. Halle: Niemeyer. Republish: 1913.

Iris, M. A., Litowitz, B. E., and Evens, M. (1988). Problems of the Part-Whole Relation. In Evens, M., editor, *Relations Models of the Lexicon*, pages 261–288. Cambridge: Cambridge U. Press.

Johansson, I. (2004). On the Transitivity of Parthood Relations. In Hochberg, H. and Mulligan, K., editors, *Relations and Predicates*, pages 161–181. Frankfurt: Ontos.

Johnston, M. (1992). Constitution Is Not Identity. *Mind*, 101:89–105.

Keefe, R. (2000). *Theories of Vagueness*. Cambridge: Cambridge U. Press.

Kline, A. D. and Matheson, C. A. (1987). The Logical Impossibility of Collision. *Philosophy*, 62:509–515.

Kuratowski, C. (1922). Sur l'opération A^- de l'Analysis Situs. *Fundamenta Mathematicae*, 3:182–199.

Leonard, H. S. and Goodman, N. (1940). The Calculus of Individuals and Its Uses. *J. of Symbolic Logic*, 5:45–55.

Leśniewski, S (1916). *Podstawy ogólnej teoryi mnogości. I.* Moskow: Prace Polskiego Koła Naukowego w Moskwie.

Lewis, D. K. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. K. (1991). *Parts of Classes*. Oxford: Blackwell.

Li, S. and Ying, M. (2004). Generalized Region Connection Calculus. *Artificial Intelligence*, 160:1–34.

Lowe, E. J. (1989). *Kinds of Being*. Oxford: Blackwell.

Lowe, V. (1953). Professor Goodman's Concept of an Individual. *Philosophical Review*, 62:117–26.

Markosian, N. (1998a). Brutal Composition. *Philosophical Studies*, 92: 211–249.

Markosian, N. (1998b). Simples. *Australasian J. of Philosophy*, 76:213–228.

Martin, R. M. (1965). Of Time and the Null Individual. *J. of Philosophy*, 62: 723–736.

Masolo, C. and Vieu, L. (1999). Atomicity vs. Infinite Divisibility of Space. In *Proc. COSIT 99*, pages 235–250. Berlin: Springer.

McGee, V. (1997). "Kilimanjaro". *Canadian J. of Philosophy*, 23:141–195. Suppl.

McGrath, M. (1998). Van Inwagen's Critique of Universalism. *Analysis*, 58: 116–121.

Merricks, T. (1999). Composition as Identity, Mereological Essentialism, and Counterpart Theory. *Australasian J. of Philosophy*, 77:192–195.

Moltmann, F. (1997). *Parts and Wholes in Semantics*. Oxford: Oxford U. Press.

Morison, B. (2002). *On Location: Aristotle's Concept of Place*. Oxford: Clarendon.

Mukerjee, A. and Joe, G. (1990). A Qualitative Model for Space. In *Proc. AAAI 90*, pages 721–707. Menlo Park: AAAI Press.

Needham, P. (1981). Temporal Intervals and Temporal Order. *Logique et Analyse*, 24:49–64.

Noonan, H. W. (1993). Constitution Is Identity. *Mind*, 102:133–146.

Nutt, W. (1999). On the Translation of Qualitative Spatial Reasoning into Modal Logics. In *Proc. KI 99*, pages 113–124. Berlin: Springer.

Parsons, J. (2000). Must a 4-Dimensionalist Believe in Temporal Parts? *The Monist*, 83:399–418.

Parsons, J. (2004). Distributional Properties. In Jackson, F. and Priest, G., editors, *Lewisian Themes*, pages 173–180. Oxford: Clarendon.

Parsons, J. (2006). Theories of Location. *Oxford Studies in Metaphysics*, 3:201–232.

Parsons, T. and Woodruff, P. (1995). Worldly Indeterminacy of Identity. In *Proc. Aristotelian Society*, page 95. 171–191.

Peirce, C. S. (1983). The Logic of Quantity. Published posthumously in Hartshorne, C. and Weiss, P., editors *Collected Papers of Charles Sanders Peirce*, Vol. IV, pages, 85–152. Cambridge MA: Harvard U. Press, 1933.

Perzanowski, J. (1993). Locative Ontology. *Logic & Logical Philosophy*, 1: 7–94.

Pianesi, F. and Varzi, A. C. (1996a). Events, Topology, and Temporal Relations. *The Monist*, 78:89–116.

Pianesi, F. and Varzi, A. C. (1996b). Refining Temporal Reference in Event Structures. *Notre Dame J. of Formal Logic*, 37:71–83.

Polkowski, L. and Skowron, A. (1994). Rough Mereology. In *Proc. ISMIS 94*, pages 85–94. Berlin: Springer.

Pratt, I. and Schoop, D. (2000). Expressivity in Polygonal, Plane Mereotopology. *J. of Symbolic Logic*, 65:822–838.

Pyle, A. (1995). *Atomism and Its Critics*. Bristol: Thoemmes.

Randell, D. A., Cui, Z., and Cohn, A. G. (1992). A Spatial Logic Based on Regions and Connections. In *Proc. KR 92*, pages 165–176. Los Altos CA: Morgan Kaufmann.

Rea, M., editor (1997). *Material Constitution*. Lanham MD: Rowman & Littlefield.

Rea, M. (1998). In Defense of Mereological Universalism. *Philosophy and Phenomenological Research*, 58:347–360.

Renz, J. (1998). A Canonical Model of the Region Connection Calculus. In *Proc. KR 98*, pages 330–341. San Francisco: Morgan Kaufmann.

Rescher, N. (1955). Axioms for the Part Relation. *Philosophical Studies*, 6:8–11.

Ridder, L. (2002). *Mereologie*. Frankfurt: Klostermann.

Roeper, P. (1997). Region-Based Topology. *J. of Philosophical Logic*, 26: 251–309.

Rosen, G. and Dorr, C. (2002). Composition as a Fiction. In Gale, R., editor, *The Blackwell Guide to Metaphysics*, pages 151–174. Oxford: Blackwell.

Sanford, D. (1993). The Problem of the Many, Many Composition Questions, and Naive Mereology. *Noûs*, 27:219–228.

Schuldenfrei, R. (1969). Eberle on Nominalism in Non-Atomic Systems. *Noûs*, 3:427–430.

Sharvy, R. (1983). Mixtures. *Philosophy and Phenomenological Research*, 44:227–239.

Shorter, J. M. (1977). On Coinciding in Space and Time. *Philosophy*, 52: 399–408.

Sider, T. (1993). Van Inwagen and the Possibility of Gunk. *Analysis*, 53: 285–289.

Sider, T. (2001). *Four-Dimensionalism*. Oxford: Clarendon.

Simons, P. M. (1987). *Parts*. Oxford: Clarendon.

Simons, P. M. (1991a). Faces, Boundaries, and Thin Layers. In Martinich, A. P. and White, M. J., editors, *Certainty and Surface in Epistemology and Philosophical Method. Essays in Honor of Avrum Stroll*, pages 87–99. Lewiston: Edwin Mellen Press.

Simons, P. M. (1991b). Free Part-Whole Theory. In Lambert, K., editor, *Philosophical Applications of Free Logic*, pages 285–306. Oxford: Oxford U. Press.

Simons, P. M. (1991c). Part/Whole II: Mereology Since 1900. In Burkhardt, H. and Smith, B., editors, *Handbook of Metaphysics and Ontology*, pages 209–210. Munich: Philosophia.

Simons, P. M. (2003). The Universe. *Ratio*, 16:237–250.

Simons, P. M. (2004). Extended Simples: A Third Way between Atoms and Gunk. *The Monist*, 87:371–384.

Smith, B. (1994). *Austrian Philosophy*. La Salle IL: Open Court.

Smith, B. (1995). On Drawing Lines on a Map. In *Proc. COSIT 95*, pages 475–484. Berlin: Springer.

Smith, B. (1996). Mereotopology: A Theory of Parts and Boundaries. *Data & Knowledge Engineering*, 20:287–304.

Smith, B. (1997). Boundaries: An Essay in Mereotopology. In Hahn, L.H., editor, *The Philosophy of Roderick Chisholm*, pages 534–561. La Salle IL: Open Court.

Smith, B. (2001). Fiat Objects. *Topoi*, 20:131–148.

Smith, B. and Mark, D. (2003). Do Mountains Exist? Towards an Ontology of Landforms. *Environment & Planning B*, 30:411–427.

Smith, B. and Varzi, A.C. (2000). Fiat and Bona Fide Boundaries. *Philosophy and Phenomenological Research*, 60:401–420.

Sorabji, R. (1988). *Matter, Space, and Motion*. Ithaca NY: Cornell U. Press.

Stell, J. G. (2000). Boolean Connection Algebras: A New Approach to the Region-Connection Calculus. *Artificial Intelligence*, 122:111–136.

Stell, J. G. and Worboys, M. F. (1997). The Algebraic Structure of Sets of Regions. In *Proc. COSIT 97*, pages 164–174. Berlin: Springer.

Strawson, P. F. (1959). *Individuals*. London: Methuen.

Stroll, A. (1988). *Surfaces*. Minneapolis: U. of Minnesota Press.

Tarski, A. (1929). Les fondements de la géométrie des corps. *Annales de la Société Polonaise de Mathématique*, 7:29–33. suppl.

Tarski, A. (1935). Zur Grundlegung der Booleschen Algebra. I. *Fundamenta Mathematicae*, 24:177–198.

Thomson, J. J. (1983). Parthood and Identity Across Time. *J. of Philosophy*, 80:201–220.

Thomson, J. J. (1998). The Statue and the Clay. *Noûs*, 32:149–173.

Tiles, J. E. (1981). *Things That Happen*. Aberdeen: Aberdeen U. Press.

Tsai, H-C. (2005). *The Logic and Metaphysics of Part-Whole Relations*. Phd dissertation, Columbia U.: Department of Philosophy.

Tversky, B. (1989). Parts, Partonomies, and Taxonomies. *Developmental Psychology*, 25:983–995.

van Benthem, J. (1983). *The Logic of Time*. Dordrecht: Reidel. Republish: 1991.

van Cleve, J. (1986). Mereological Essentialism, Mereological Conjunctivism, and Identity Through Time. *Midwest Studies in Philosophy*, 11:141–156.

van der Zee, E. and Slack, J., editors (2003). *Representing Direction in Language and Space*. Oxford: Oxford U. Press.

van Inwagen, P. (1981). The Doctrine of Arbitrary Undetached Parts. *Pacific Philosophical Quarterly*, 62:123–137.

van Inwagen, P. (1990). *Material Beings*. Ithaca NY: Cornell U. Press.

van Inwagen, P. (1993). Naive Mereology, Admissible Valuations, and Other Matters. *Noûs*, 27:229–234.

van Inwagen, P. (1994). Composition as Identity. *Philosophical Perspectives*, 8:207–220.

Varzi, A. C. (1994). On the Boundary Between Mereology and Topology. In *Proc. 16th Wittgenstein Symposium*, pages 423–442. Vienna: HPT.

Varzi, A. C. (1996a). Parts, Wholes, and Part-Whole Relations: The Prospects of Mereotopology. *Data & Knowledge Engineering*, 20:259–286.

Varzi, A. C. (1996b). Reasoning about Space: The Hole Story. *Logic & Logical Philosophy*, 4:3–39.

Varzi, A. C. (1997). Boundaries, Continuity, and Contact. *Noûs*, 31:26–58.

Varzi, A. C. (2000). Mereological Commitments, *Dialectica*, 54:283–305.

Varzi, A. C. (2001). Vagueness in Geography. *Philosophy & Geography*, 4: 49–65.

Varzi, A. C. (2003). Mereology. In Zalta, E. N., editor, Stanford Encyclopedia of Philosophy. Stanford: CSLI. on line publication.

Varzi, A. C. (2005). A Note on the Transitivity of Parthood. *Applied Ontology*, 1:141-146.

Varzi, A. C. (2006). The Universe among Other Things. *Ratio*, 19:107–120.

Whitehead, A. N. (1919). *An Enquiry Concerning the Principles of Human Knowledge*. Cambridge: Cambridge U. Press.

Whitehead, A. N. (1920). *The Concept of Nature*. Cambridge: Cambridge U. Press.

Whitehead, A. N. (1929). *Process and Reality*. New York: Macmillan.

Wiggins, D. (1968). On Being in the Same Place at the Same Time. *Philosophical Review*, 77:90–95.

Wiggins, D. (1980). *Sameness and Substance*. Oxford: Blackwell.

Winston, M., Chaffin, R., and Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11:417–444.

Yi, B.-U. (1999). Is Mereology Ontologically Innocent? *Philosophical Studies*, 93:141–160.

Yoes, M. (1967). Nominalism and Non-Atomic Systems. *Noûs*, 1:193–200.
Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, 8:338–353.
Zimmerman, D. W. (1996a). Could Extended Objects Be Made Out of Simple Parts? *Philosophy and Phenomenological Research*, 56:1–29.
Zimmerman, D. W. (1996b). Indivisible Parts and Extended Objects. *The Monist*, 79:148–180.

Index

Abscissa, 369
Absolute, 613, 615, 623, 625–626, 630, 638, 648, 651–652, 656
  distance, 176
  time, 612–613, 623, 625
  value, 619
Accessibility relation, 108, 144, 150, 400
Addition in affine plane, 371
Adhesion, 977
Adjoint, 387
Adjunction, 885
  between toposes, 470
  on sets, 888
Affine
  bisimulation, 278
  equation, 377
  geometry, 228, 344
  mapping, 640
  operation, 653
  operations, 365
  plane over ternary ring, 371
  plane, 365
  relation, 381
  space, 1, 364–365
  structure, 364
  subgeometry, 366
  transformation, 344, 378
Affinity, 378
After, 656
Alexander's horned sphere, 71
Alexandroff, 724
  extension, 252
  space, 238, 508, 570
  spaces and Kripke frames, 508, 570
  topology, 306, 890
Alexandroff-Zeeman theorem, 641, 651
Algebra of binary relations, 111
  full, 111
Algebraic, 100, 449
  logic, 104, 111, 653–654, 659
Almost fixed point, 721
Alternate sequential filter, 892
Anti-granulometry, 896
Anti-matroid, 782
Approximation
  algorithms, 180
  of a semi-algebraic set, 831, 847
Aristotle
  Physics, 713
Arithmetic universe, 493
Arrow
  logic, 283, 289
  model, 289
Artificial Intelligence, 7, 9, 93, 102, 119, 134
Associated sheaf functor, 483
Asymptotically flat, 688, 697, 699, 703
Atom
  extended (mereological), 1021
  mereological, 961
  unextended (mereological), 963
Atomic
  permanence, 312
  relations, 165
Atomicity
  mereological, 962
  of space, 1017, 963
Atomism, 961
Atomless algebra, 125
Atomless gunk, 961
Atomlessness
  mereological, 962
  of space, 1017
Attribute geometric, 801
Automated reasoning, 102, 136
Automorphism, 879
  dual, 879
Axiom of social geometry, 10
Backtracking, 164, 168
  heuristics, 168, 194
  strategies, 195
Ball, 32
Base relations, 165
Basic relations, 165
Basis, 231, 360
Beth's theorem, 649–650
Betweenness, 228, 349, 386, 655
  relation, 386
  structure, 386

Bi-Heyting, 740
Bi-persistent, 310
Big Bang, 661, 691, 696–697
Big Crunch, 691
Bimodal dynamic topological logic of homeomorphisms, 586
Binary relation, 110, 164
Bipath-consistency algorithm, 201
Bisimulation, 5
  affine, 278
  topological, 225
Bivalence, 975
Black hole, 607, 660–661, 664–665, 676, 684–687, 691–692, 704
  electrically charged, 697, 699, 702, 705
  Reissner-Nordström, 697
  rotating, 697, 699–702, 705
  Schwarzschild, 684, 686–687, 689, 694, 696–698, 700, 703
Block, 761
  algebra, 175
Body, 618
Boolean algebra, 14, 105, 136, 654, 968
  atom, 105
  dense, 105
  with operators, 106
    additive, 107
    normal, 107, 109
  atomic, 105, 112
  canonical extension, 106
  complete, 105, 112
  regular subalgebra, 106
Boolean contact algebra, 122, 132
Boolean matrix, 53
Boundary, 113, 232, 516, 741, 744, 956, 985, 990, 995
  and common sense, 991
  and physics, 978
  as a conceptual idealization, 991
  as lower-dimensional, 996
  as ontologically dependent, 995
  coincidence, 1020, 995
  dismissed, 996
  does not occupy space, 1020, 995
  fiat, 994–995
  mereological, 971
  ownership of, 993
  Part, 995
  potential, 994
Boundarylessness, 1005, 984
  implies atomlessness, 998
  of space, 1017
Bounded morphism, 402
Boundedness
  definability of, 30, 57, 74
  mereological, 965
Box, 401
Branching factor
  average, 168, 192
Bruhat order, 794
CAD, 816–818, 826, 828
Candidate sets, 190
Canonical
  frame, 109
  model, 402
  relation, 109
  topo-model, 239
  topological space, 239
Cantor-Bendixson derivative, 332
Cardinal
  algebra, 174
  directions, 173
Carrier, 448
  empty, 448, 451
Cartesian
  category, 450, 458, 461
  functor, 450
  space, 101
  theory, 449
Categorical, 639
  logic, 457
Category
  ∞-positive, 461
  Cartesian, 450, 458, 461
  geometric, 461
  regular, 461
  theory, 457
Causal past, 700
Causal-geometry, 653, 657
Causal-relation, 653–656
Causality relation, 656
Causally precedes, 656
Causally separated, 652
Cavity, 1008–1009
  donut-shaped, 1009
Cayley graph, 792
Cell, 19, 816
  decomposition theorem, 19, 816
  finite cell decomposition theorem, 817
  geometry, 752
  partition, 759
Change, 531, 958
  mereological, 1013, 1022, 1027, 958
  of location, 1013, 1027
  topological, 1013, 992
Check-Refinements algorithm, 187
Clan, 128
  saturated, 130
Classification theorem, 71
Classifying topos, 461, 466
Clause constraints, 182
Clopen, 232
Closed
  entity, 991

set, 231, 301
timelike curve, 701
Closer-than relation, 394,
Closer operator, 524
Closing, 286
activity ordering, 895
algebraic, 880
area, 866
binary, 862
by reconstruction, 865
flat, 870
grey-level, 874
structural, 882
topological, 881
Closure, 232, 301
algebra, 108–109, 116, 140–141, 144
mereotopological, 990
of relations, 185
operator, 108, 113, 116, 510
space, 723
system, 725
Cluster, 128, 130, 132–133, 241
proximity (p-cluster), 131
Co-derivative, 232
Co-Heyting, 740
Co-location, 1015, 1019–1020, 1022, 986, 996
and interpenetration, 1020
and mirroring principles, 1019
Co-planar lines, 362
Co-punctuality, 349, 359
Cocompletion, 480
Coequalizer, 454
Coherent
formula, 436
theory, 436, 446
type theory, 454
Cohesion, 977
Coincidence, 1020, 951, 995
Colimit, 463
filtered, 464, 470
Collinear, 651–652, 655
Collinearity, 349, 359
Collineation, 361, 641
group, 361
Combination of strategies, 195
Comma square, 470
Common knowledge, 272, 327
Compact, 233
Compactification
of mereotopology, 34
of topological space, 32
Compactness, 491
Companions, 260
Compatibility axiom, 124
Complement, 135, 164
mereological, 968
vs. mereological complement, 999
Complementation
mereological, 1006, 960, 964, 988
Complete lattice, 876
Complete, 639, 646–647, 659, 681
Completely prime filter, 438
Completeness, 169
homeomorphisms on the reals, 586
topological, 38
with respect to a theory, 86
Complex algebra, 109
Complex, 822
Complexity classes
P, 178
NP, 178
Composition, 110, 112, 164, 289, 457, 639, 654
as identity (mereological), 970
horizontal, 463
mereological, 964, 969
ontologically extravagant (mereological), 970
ontologically exuberant (mereological), 970
table, 112, 122, 139, 166
vagueness of (mereological), 969
Compositional reasoning, 138
Computational completeness
of a query language, 849, 851
Computational complexity, 150, 178
Conceptual analysis, 612, 638, 659–660
Concurrence, 349
Cone, 819, 837
radius, 819, 837, 839
type, 819, 821
Conformal diagram, 697–699
Conical structure
local, 819, 837
Conjunction, 458
rule, 436
Connected parthood, 983
Connected, 113, 222, 233, 511
element, 39
relation, 170
Connectedness, 518
definability of, 27
firm, 1003
self-, 1003, 986, 988
simple, 1007
Connection, 101, 119
n-, 979
algebra, 738
and location theory, 1028, 986, 996
and physics, 978
and ‘contact’, 977
and ‘touching’, 977
as adhesion, 977
as cohesion, 977
axiom, 124, 130
Connection
Axioms for

Additivity (C.11), 990
Boundarylessness (C.5), 984
Converse Monotonicity (C.4), 982
Dependence (C.12), 995
Fusion Connection (C.8), 989
Idempotence (C.10), 990
Inclusion (C.9), 990
Integrity (C.3$_a$), 981
Left Join (C.6), 986
Monotonicity (C.3), 981
Open Density (C.14), 1005
Reflexivity (C.1), 979
Restricted Boundarylessness (C.5$_\phi$), 997
Right Join (C.7), 986
Symmetry (C.2), 979
Topological Fusion (C.13), 999
Unity (C.3$_b$), 981
basic principles, 978
bridging principles, 980
by coincidence, 994
By-, 979
complete, 1003
definability of, 976, 986
extensional, 980, 985
extensive, 956
external, 1002, 981
firm, 1003
in point-set topology, 990
interpretation of, 1001, 1028, 982, 990
Kuratowski Axioms for, 990
loose, 1004
Mediate, 979
monotonicity of, 981–982
perfect, 1003
reflexivity of, 979
relation, 135, 751
restricted, 1029
strength of, 1002, 978
sufficient for fusion, 987
symmetry of, 979
theory, 715, 736
topological vs. metric, 1004, 977
transitive, 980
varieties of, 1002
vs. overlap, 980, 997
Connectivity
graph, 829
topological, 831, 845–846
Consistency, 164, 505
problem, 167
Consistent, 639–640, 644, 647
refinement, 167
scenario, 167
set of formulas, 239
Constraint satisfaction problem, 163
finite domains, 164
infinite domains, 164
solution, 164
Constraint, 9, 163
language, 172
linear, 803, 805, 809–810
network, 163
polynomial, 803, 805, 807
propagation, 164
Contact relation, 22, 104, 118–120, 122, 127, 171, 977
algebra, 83, 121, 146
definability of, 38
standard, 120, 129, 134
Context, 446
weakening, 447, 458
Continuity, 477
axiom, 397
schema, 391, 397,
of change, 531
Continuous, 233
flat functor, 479
function, 550, 566
image, 233
ordering, 387
Contractible, 720
Converse, 110, 289
Conversion, 164
Convex, 250, 755
closure, 281
Convexity structure, 749
Coordinate frame
global, 664–665, 667, 672, 674
Coordinate system, 610, 618, 646, 648, 664–665
Cartesian, 610
Coordinate-transformation, 639
Coordinates, 368
ring, 370
set, 368
system, 368
Coproduct, 454
disjoint, 460
Correctness, 169
Correspondence
principle, 721
theory, 403
Cosmic Censor Hypothesis, 661
Cosmology, 607
Coverage, 479
theorem, 491
Coxeter matroid, 794
Crosing lines, 381
Cross Axiom, 312
Cross-cut, 35
Crosspolytope, 769
CSPENT, 169
CSPMIN, 169
CSPSAT, 167, 185–186
CTC, 701, 703

Cumulativity, 1017
Curvature, 608, 668, 678, 692, 701, 703, 705
Curve, 668, 677, 689
  photonlike, 671, 678
  smooth, 668
  time-faithful, 669–670, 678, 690, 700
  timelike, 668, 670, 688
Curve-selection, 35
  theorem, 36, 815
Curved, 692, 696
Cut
  rule, 436
  transformation, 833–836
Cycle law, 113
Cycle, 720
Cylindric algebra, 108, 654
Cylindrical algebraic decomposition, 816–818, 826, 828
  sample points, 826
  sign conditions, 826
  basis phase, 827
  extension phase, 827
  projection phase, 827
  sector, 827
  stack, 827
Dacey, 734
Data
  spatio-temporal, 813
Database, 7, 9
  constraint, 799, 803, 807, 812
  polynomial constraint, 807
  relation, 799
  relational, 799, 810, 842
  isotopic, 833
Decidability, 304, 315, 322, 505
Decidable, 639, 646–647
Decision
  problem, 178, 505
  procedure, 150
Decomposition
  cylindrical algebraic, 816–818, 828
  finite cell, 817
  topological, 815–816
Dedekind
  complete, 387
  section, 455, 471, 473
Definability theory, 648–650, 657–658, 704
Definite description, 134
Definite knowledge, 165
Definitional equivalence, 648–650, 653, 658, 681
Definitional expansion, 649–651, 704
Deletion subgeometry, 366
Dense, 129, 232, 387, 511
Dense-in-itself, 232
Density
  mereological, 964
  of space, 1017
  open, 1005
Dependent lines, 357
Derivative, 232
Derived set, 332
Desargues property, 391
  first, 372
  second, 374
  third, 374
Desarguesian plane, 376
Deterministic algorithm, 178
DIA, 204
Diamond, 401
Dichotomy theorem, 829–831
Difference modality, 306, 336
Difference
  mereological, 954, 960, 968
Differential, 662, 674–675
Digital
  n-cube, 718, 748
  n-sphere, 747
  line, 256
  plane, 256
Dilation, 286, 379
  commutativity, 380
  transitivity, 380
  algebraic, 886
  binary, 861
  by a relation, 888
  flat, 870
  fuzzy, 909
  grey-level, 872
  logics, 920
  possibility modality, 925
Dimension, 743
  of linear space, 360
  of orthogonality frame, 355
Dimensionality, 150
Direct image, 459, 462
Directed, 439
  intervals, 204
    algebra, 204
  join, 439, 464
  space, 313
Direction relation matrix, 176
Disc, 120, 122
Discontinuity, 131
Discrete, 232
  space, 9, 151
Discriminator algebra, 111
Disjoint
  coproduct, 460
Disjointness
  mereological, 953
Disjunction, 164, 459
  rule, 436
Disjunctive relations, 164
Dismantlable, 720

Dissectiveness, 1017
Dissectivity number, 1007
Distance, 176, 507, 622, 628, 638
  Euclidean, 508, 640–641
  Minkowski, 638, 656, 669, 672, 681
  radar, 692, 695
  relativistic, 638, 640–641, 648, 651, 656, 658, 669, 672, 681
  spatial, 621, 624
  time, 621, 624
Distance space, 507
Distributive
  law, 730
Distributivity
  frame, 432, 437
Division ring, 350
  strong, 371
Donut, 1007
  punctured, 1008
Double cross calculus, 174
Double powerlocale, 491
Downset, 238
DSO space, 334
Dual matroid, 766
Dual p-algebra, 739
Dual universe, 697–698
Duality, 639, 648, 658–659
  in projective planes, 363
  by complementation, 861
  by grey-level inversion, 870
  by lattice inversion, 880
  lattice-theoretical, 879
Dynamic Kripke
  frame, 572
  model, 572
Dynamic metric logic, 556
Dynamic metric model, 556
Dynamic topological logic, 550, 566, 571
  a decidable example, 574
  Alexandroff spaces, 573
  and Poincaré Recurrence, 573
  completeness results for fragments, 568
  fragments, 573
  homeomorphisms, 586
  next-interior fragments, 573
  nonaxiomatizability results, 567
  purely temporal fragments, 573, 576–577
  purely topological fragments, 573, 576
Dynamic topological model, 550, 566, 571
Dynamic topological system, 566, 571
  bisimulation, 597
Eddington-Finkelstein re-coordinatization, 688–689, 693, 699
EF-game, 5–6, 829, 841
Effective finite model property, 241
Efficiency, 178
Einstein's
  cosmological term, 701, 705
  field equations, 704–705
  light-clock, 634–635
Electrodynamics, 704
Element
  generalized, 459
  generic, 459
  global, 459
Elementarily
  embeddable, 22
  equivalent, 22
Elementary
  submodel, 22
  topos, 465
Embedding in space, 1030
Empty
  locale, 440
  relation, 164
Enclosure (topological), 980
  vs. parthood, 982, 999
End-cut, 35
Entailment problem, 169
Entity
  closed, 991
  co-located, 1020
  disconnected, 987
  four-dimensional, 1014, 1022, 958
  geographic, 978
  immaterial, 1020
  located in space, 945
  null, 952, 966
  open, 991
    vs. Open*, 1000
  particular vs. universal, 1015
  reference, 1025
  scattered, 961, 969, 975, 987
  spatial, 945
  target, 1025
  universal, 952
  vague, 972
Equal segments construction, 396
Equality, 4, 458, 952
  and identity, 953
  rule, 447
Equation of line, 369
Equational
  class, 107–108, 117
  reasoning, 140
Equidistance, 349, 525, 655
Erlangen program(me), 5, 344
Erosion, 286
  algebraic, 886
  binary, 861
  by a relation, 888
  flat, 870
  fuzzy, 909
  grey-level, 872

logics, 920
necessity modality, 925
Essentialism
mereotopological, 1013
Essentially algebraic, 450
Essentially propositional, 471
Etale, 479
Euclid, 2, 91
fifth postulate, 344, 393, 397
Euclidean
dilation, 641
field, 351
geometry, 4, 101, 150, 629, 668, 680
hierarchy, 255
isometry, 640
plane, 102, 120
Euler
formula, 47
Event horizon, 664, 687, 691, 695, 698, 700–701, 703
Event, 611, 622, 624, 627, 679
Exchange property, 360
Exhaustive split, 168
Existential
commitment, 102
rule, 447
Explicit definition, 650
Exponential, 465
Expressive power, 150, 531
Expressiveness
first-order, 829
on finite relations over the reals, 829
on infinite relations, 829
Expressivity and ontology, 1001
Extension
proper, 952
interior, 982
tangential, 982
simple functional, 484
Extensionality
axiom, 83, 119–121, 124, 129, 135, 151
mereological, 1020, 1022, 957–959, 967–968, 972, 985
and atomism, 962
and supplementation, 957
and vagueness, 972, 975
topological, 980, 985, 989
Extensive connection, 22, 91
Extent, 440
Exterior
mereotopological, 990
Extremally disconnected, 233
Face, 750, 755
Fano
axiom, 385
plane, 364
Fibre, 490
Fibred
locale, 490
product, 463, 475
Field, 351, 619–620, 629, 638, 667, 672, 680
real numbers, 805
Euclidean, 619
finite, 620
linearly ordered, 618–619
ordered, 619, 676
quadratic, 640–641, 643–647, 650–651, 664, 681–682
quadratic, 619, 626, 629, 639
real-closed, 647, 682
Filter
completely prime, 438
Filtered colimit, 464, 470
Filtering functor, 480
Finite model property, 235
Finite structures over the reals, 830
Finite-cofinite algebra, 124
Finitely decomposable
mereotopology, 20
structure, 39
Finiteness
Kuratowski, 454
First response, 196
First-order logic, 1, 5–6, 8, 21, 522, 608, 639, 646–647, 649–651, 658, 660, 662, 676–677, 683, 690
over the reals, 805
with linear constraints, 805, 809, 811, 824
with operators, 842
with polynomial constraints, 805, 807–808, 825
with topological operators, 851–852
with transitive closure, 842
with while-loop, 841, 849,
First-order temporal logic, 557
First-order theory, 22, 171
Five-segment axiom, 397
Flat functor, 480
continuous, 479
Flat oriented matroid, 783
Flat zone, 871
Flat, 696
Formal concept analysis, 897
Formal semantics of relations, 169
Formal topology, 438, 491
Formula
coherent, 436
first-order, 805
geometric, 436, 446
in context, 446
order-generic, 830
quantifier-free, 805–806, 810, 813
equivalence, 824
prenex formal form, 826
Four-dimensionalism, 1014, 1022, 958

Frame of reference, 173
Frame, 108, 234, 305, 432, 436, 485, 490
  distributivity, 432, 437
Free
  algebra, 466
  model, 451, 454, 461–463, 466
Frobenius rule, 447, 461
Full infinite binary tree, 242
Full infinite quaternary tree, 235
Function, 445
Functor
  Cartesian, 450
Fusion, 235, 954, 965, 967
  and vagueness, 969
  implausible, 969
  infinitary, 966
  over and above the parts, 970
  unique, 967
  unrestricted, 967, 969
  vs. mereological fusion, 999
  vs. set abstraction, 968
  vs. upper bound, 965
Galaxy, 659
Gale order, 791
Galois
  connection, 884
    on sets, 885
  map, 884,
Game of Life, 548
General Principle of Relativity, 664
Generalized
  element, 459
  model, 434
  point, 439, 469
  space, 429, 487
Generic
  collapse, 829–830
  element, 459
  model, 434, 468
  point, 439
Genus (topological), 1007
Geodesic, 662, 668–669, 674, 680, 683, 689–690, 703, 705
  photonlike, 671, 674, 677, 679, 682, 684, 690, 697
  timelike, 670–671, 674, 677, 679, 682, 684, 690, 692, 702
Geodesical reconstruction, 865
Geographic information system, 512, 810–811
Geometric
  attribute, 801
  category, 461
  formula, 436, 446
  morphism, 462, 469
  theory, 436, 446, 484
  type theory, 452–453, 461, 484
  type, 463
Geometrization, 608, 638
Geometry of Solids, 3–4, 6, 92, 102
Geometry, 657, 664, 682
  absolute, 345
  affine, 228, 344
  elementary, 345
  Euclidean, 4, 101, 150, 629, 668, 680
  general affine, 391
  hyperbolic, 344
  Minkowski, 638, 648, 650
  o-minimal, 812
  observer-independent, 648
  of Solids, 102
  projective, 345, 765
  real algebraic, 812
Get-Refinements algorithm, 187
Giraud's theorem, 461, 485
GIS, 512, 810–811
Global
  element, 459
  point, 440, 467
Glue transformation, 833–836
Grain, 864
Granularity, 165
  mereological, 963
Granulometry, 896
Graph
  convexity, 726
  neighbourhood, 43
  Dacey, 734
Gravity, 608, 660–664, 668, 674, 686, 690, 692, 700
Greedy algorithm, 791
Grey-level images, 867
  ordering, 867
  supremum and infimum, 867
Grothendieck
  topology, 479
  topos, 460
Group, 350, 643
Gunk
  atomless, 961
H-open, 263
Half-space, 20
Hall Marshall Jr.
  planar ternary ring, 370
Hausdorff space, 233
Heine-Borel theorem, 443, 445
Helly graph, 721
Hereditarily irresolvable, 253
Heuristics, 180
  backtracking, 168, 194
Heyting algebra, 117, 432, 740
  pseudo-complement, 117
Hilbert
  10th problem, 526, 832
  axiomatization, 2
  axioms for incidence, 359

coordinatization, 651
system, 59
Hit-or-miss transform, 861
Hole, 1003, 1006, 1021, 985, 992
and ontology, 1010
as a negative object, 1010
internal, 1008
knotted, 1009
perforating, 1008
Homogeneous
mereotopology, 56
Homogeneously embedded, 56
Homomorphism
frame, 432, 490
model, 448–449
Homothety, 379
Horizontal topology, 263
Horizontally open, 263
HORNSAT, 179
HV-open, 263
Hyper-rectangular convex, 254
Hyperplane
arrangement, 774
axis-oriented, 58
rational, 21
Hyperspace, 490
Idempotence, 863
Identity, 878, 953, 958
contingent, 958
diachronic, 958
refinement, 190
relation, 110, 164
rule, 436
through change, 958
transitivity of, 1022
Image, 458, 460, 463
Imaginary numbers, 619
Implication, 459
Implicit definition, 649–650
Implicit knowledge, 273
Incidence, 349, 358
basis, 359
frame, 358, 413
affine, 414
projective, 414
geometry, 359
Indefinite knowledge, 165
Independent, 639, 645
lines, 357
points, 360
linearly, 667
Indiscernibility, 719, 954
Induced subgraph, 719
Induction algebra, 452
Inertial, 612
body, 620
motion, 609
reference frame, 609
Infinitely chequered subsets of $\mathbb{R}^\infty$
logic of, 255
Initial model, 450, 461
Instance
database, 807, 810
relation, 810
Interaction of relations, 198
Interchange law, 463
Interdefinability, 102
Interface (topological), 1011
Interior, 232–233, 301, 724
algebra, 108
image, 233
mereotopological, 990
operator, 108, 113, 116
point, 232
vs. mereotopological interior, 1000
Interpolation
axiom, 83, 124, 130, 133–134
Intersection, 164, 349, 359
Interval algebra, 166
Intractability, 178
Inverse, 619
image, 457, 462
Involuted monoid, 113
Involution, 758, 769
Irreflexive fragment, 258
Irresolvable, 253
Isomorphism, 5, 643, 649–651, 653, 674–676, 681, 684–685, 693, 879
between linear spaces, 361
dual, 879
Isotopy, 832
Isotropic, 609, 621, 660
JEPD relations, 165
Join
directed, 439, 464
Jointly exhaustive relations, 165
Jordan arc, 35
Jordan curve, 35, 136
theorem, 35
converse of, 36
KD45, 329
Kennison's theorem, 451
Khalimsky
line, 256
plane, 256
Kinematics, 992
Newtonian, 608, 610
special relativistic, 608, 610, 612
Knowledge Representation, 102, 119, 134
Kripke
frame, 108, 400, 570
logic of a class of, 401
model, 401
semantics, 143, 305

Kruskal's tree theorem, 556
Kuratowski
axioms, 507
finiteness, 454
theorem, 46
Lambek calculus with permutation, 287
Lattice
inversion, 880
Law of Nature, 663–664
Left exact
sketch, 450
theory, 450
Leibniz's law, 958–959
Leibniz's principle, 646
Length, 746
Length-contraction, 632, 634–637, 640
Liar paradox, 661, 701
Lifeline, 611, 620
Light Axiom, 609, 612, 616, 621, 623, 642, 645, 660, 663, 680
Light signals, 608
Light-cone, 642, 652–654, 668, 688, 691, 695, 698–699, 703
Lightlike separated, 653
Limit point, 232
Lindenbaum algebra, 143–144, 433, 439, 451, 461, 466
Line of slopes, 369
Line-element, 673, 684, 686, 695, 697–698, 702
Linear mapping, 639
Linear Time Temporal Logic, 527, 566, 577
Linear
algebra, 285
space, 359–360
planar, 362
transformation, 361
variety, 360
Lines, 349, 358
List type, 454
Local
conical structure, 819
frame, 665, 667–671
homeomorphism, 475–476, 478–479
inertial frame, 683
section, 476
Locale, 438, 471, 486
empty, 440
one-point, 440
Localic
reflection, 487
topos, 471
Locally Special Relativity Principle, 663–665
Location, 1012, 946
ψ-, 1018
as connection of a kind, 1028
Location
Axioms for
Bottom Mirroring (L.13), 1019
Conditional Emptiness (L.5), 1016
Conditional Fullness (L.17), 1021
Conditional Reflexivity (L.4), 1015
Conditional Spatiality (L.15), 1019
Cumulativity (L.8), 1017
Dissectiveness (L.7), 1017
Exactness (L.2), 1015
Exclusiveness (L.16), 1021
Fullness (L.6), 1016
Functionality (L.1), 1014
R-Atomicity (L.12), 1017
R-Atomlessness (L.10), 1017
R-Boundarylessness (L.11), 1017
R-Density (L.9), 1017
Spatial Connection (L.19), 1027
Spatial Disjointness (L.18), 1027
Spatiality (L.3), 1015
Top Mirroring (L.14), 1019
basic principles, 1014
entire, 1013
interior, 1018
tangential, 1018
exact, 1013
frame of reference, 1025
generalized, 1023
generic, 1013
implies connection, 1028
implies disjointness, 1027
mirroring principles, 1018
multiple, 1015
object-centered, 1025
observer-centered, 1025
pervasive, 1013
proper, 1016
relative, 1024–1025
theories of, 1012
ubiquitous, 1013
interior, 1018
tangential, 1018
varieties of, 1012, 1023
vs. connection, 1027
vs. parthood, 1027
whole, 1023
Loop, 350
Lorentz boost, 639
Lorentz transformation, 639–641
Lossy channel system, 555
Lower powerlocale, 491
Löwenheim-Skolem theorem, 89
Majority, 829
Malament-Hogarth computer, 703
Malament-Hogarth event, 700–701
Manifold, 69, 665, 704
Lorentz, 676, 681–682
Many-sorted logic, 134
Many-sorted, 649

Map, 435, 439, 462
Matching, 482
Mathematical morphology, 8, 10, 285, 857
  adjacency, 905
  binary, 860
  computational complexity, 875
  distances, 904
  flat, 871
  fuzzy adjacency, 915
  fuzzy directional position, 916
  fuzzy distances, 913
  fuzzy sets, 906
  grey-level, 874
  logics, 920
  modal logics, 923
Matrix, 673–674, 684, 700
Matroid, 762
Matter vs. void, 1011
Maximal tractable subset, 184
Maximally consistent, 239
Meagre space, 361
Mereology, 18, 101–102, 119, 946–947
  and nominalism, 949
  and set theory, 949, 952
  and spatial reasoning, 946
  as a Boolean algebra, 968
  as a calculus of individuals, 949
  as formal ontology, 949
  Atomistic Mereology (AX), 962
  atomistic variant of, 962
  Atomless Mereology ($\bar{A}X$), 962
  atomless variant of, 962
  Basic (M), 954
  basic principles, 950
  composition principles, 954, 964
  decomposition principles, 954
  Extensional Mereology (EM), 957
  extensional, 957
  General Extensional Mereology (GEM), 967
  history of, 947
  Minimal Mereology (MM), 956
  non-well-founded, 952
  ontological innocence of, 949
  rough, 972
Mereotopology, 18, 170, 980
  and dimensionality of space, 1005
  and Euclidean space, 1017
  and spatial embedding, 1030
  and void, 1011
  Basic (T), 980
  basic principles, 978
  boundary-free, 956, 984, 996
  Boundaryless Mereotopology ($\bar{B}X$), 998
  boundaryless variant of, 998
  bridging principles, 980
  expressivity of, 1001
  finitely decomposable, 20
  General Extensional Mereotopology ($GEMT$), 989
  Kuratowski extension of, 990
  Kuratowski General Extensional Mereotopology ($KGEMT$), 990
  layered, 1006
  Minimal Mereotopology (MT), 982
  Reductive Mereotopology (RMT), 984
  vs. kinematics, 992
Merging of objects, 991–992
Metric, 283, 506
Metric-geometry, 9, 648, 651–652, 657, 681
Metric space, 506
Metric temporal model, 544
Metric-tensor, 672, 674–676, 681–683, 686, 688, 694–695, 700
Michelson-Morley experiment, 609, 663
Minimal labels problem, 169
Minkowski
  addition, 286, 860
  circle, 655–656, 685
  geometry, 681
  metric, 650
  operations, 8
  orthogonal, 651, 655, 657
  scalar-product, 656, 672
  space, 665–666
  subtraction, 860
Mitchell-Benabou language, 466
Modal logic, 1, 8, 142, 150
  axiomatic definition, 401
  axiomatization, 402
  bi-modal spatial logic, 145
  completeness for Alexandroff spaces, 511, 590
  completeness for Cantor Space, 604
  completeness for finite spaces, 511, 590
  completeness for the class of homeomorphisms, 586
  completeness for the reals, 586
  completeness theorem, 402
  finite model property, 590
  for incidence geometries, 415
  for projective geometries, 416
  of collinearity, 406
  of elsewhere, 405
  of everywhere, 405
  of orthogonality, 408
  of parallelism and intersection, 410
  of parallelism, 407
  of qualitative distance, 406
  $S4$, 144, 510
  multi-dimensional, 150
  normal, 142, 144
Modal
  algebra, 108–109, 143
  definability, 401
  encoding, 144, 150

operator, 4, 108
Model, 99, 234, 434, 448
free, 451, 454, 461–463, 466
generalized, 434
generic, 434, 468
initial, 450, 461
standard, 434, 437, 440
Monodic fragment, 557
Moore family, 880
dual, 881
Morphism
theory, 446
Morphological filter, 891
by composition of opening and closing, 864
Motion, 608, 611–612, 617, 623, 635, 664, 670
accelerated, 620
faster-than-light, 635, 659–660
non-uniform, 620
slower-than-light, 612
uniform, 620
Multi-aspect calculi, 197
Multi-location, 1015
Multi-piece region, 151
Multiplication in affine plane, 371
Multiplicative linear logic, 287
N-chequered subsets of $I\!R^n$
logic of, 254
N-chequered, 254
N-true, 254
N-valid, 254
Natural language, 102–103
Natural numbers, 452, 454
object, 466
Natural transformation, 463
Necessity operator, 108, 509
Negation, 459
Negative direction, 204
Neighbourhood, 719
function, 888
models, 223
Newtonian kinematics, 638
Newtonian mechanics, 659
Newtonian worldview, 609, 663
Nihilism
mereological, 971
Non-deterministic algorithm, 178
Non-tangential proper part axiom, 135
NOT-ALL-THREE-SAT, 179
NP
complexity class, 178
NP-complete, 179
NP-hard, 179
Null entity, 952, 966
Null region, 119, 121, 135–136, 141
Number theory, 832–833
O-notation, 178
Object classifier, 467, 474, 487
Observation oriented, 641
Observational, 646, 658, 690, 700
Observer, 610, 618, 638, 651, 662, 667, 676
accelerated, 646, 661–664, 669, 684, 686, 692
uniformly, 684–685
inertial, 609, 646, 661–663, 668, 670, 676, 682, 684
suspended, 663, 690–691, 695, 697, 702
Observer-independent, 638, 648, 651–652, 690
Occam's razor, 682
Omega-categoricity, 88
Omega-rule, 59–60
Omitting Types Theorem, 62
ONE-IN-THREE-SAT, 179
Ontological dependence, 947, 995
Ontology, 101, 947
Open entity, 991
vs. Open*, 1000
Open, 233, 439, 487
cover, 232
image, 233
map, 475
neighbourhood, 113, 231
set, 113, 231, 301, 507
Open/closed distinction, 990–991
and figure/ground opposition, 993
vs. open*/closed* distinction, 1000
Opening, 286
activity ordering, 895
algebraic, 881
area, 866
binary, 861
by reconstruction, 865
flat, 870
grey-level, 874
structural, 882
topological, 882
Operator, 878
modal, 4
topological, 4, 510
activity ordering, 894
anti-extensive, 879
composition, 878
decreasing, 878
extensive, 879
flat, 868
idempotent, 879
increasing, 878
infimum, 878
invariance domain of, 879
ordering, 878
overpotent, 891
range of, 879
supremum, 878
translation-invariant, 879
underpotent, 891
Orbit, 546

ORD-Horn, 190
Order, 386
  collinearity geometry, 388
Ordinate, 369
Orientation, 173
Origin, 368
Orthoclosed, 729
Orthocomplement, 729
Orthogonal, 778
Orthogonality, 349, 355
  2-dimensional model, 356
  frame, 355
  planar, 355–356
  real frame, 356
  space, 720
  n-dimensional model, 357
Ortholattice, 729
Orthomodular, 734
Orthoposet, 733
Over-constrained instances, 181
Overfilter, 891
Overlap
  mereological, 952
  necessary for fusion, 969
P
  complexity class, 178
Padoa's method, 352
Pairs of relations, 199
Pairwise disjoint relations, 165
Pappian plane, 376
Pappus property, 406
  first, 375
  second, 375
Paradigmatic effects, 624, 631, 641–642, 645, 659–660
Parallelism, 352, 355
  class, 353
  frame, 353
  model, 353
  postulate, 344
  pre-model, 353
  quasi, 361
  real model, 354
  strict, 349, 353, 361
  weak, 349, 353, 362
Parallelogram relation, 374
Parity, 829
Part relation, 120
Part
  ϕ-, 951
  arbitrarily demarcated, 948
  cognitively salient, 948
  direct, 951
  distinguished, 951
  extended, 948
  functional, 951
  gerrymandered, 948
  homogeneous, 948
  immaterial, 948
  material, 948
  proper, 950, 982
    interior, 1005, 1029, 982
    tangential, 1005, 982
  self-connected, 948
  spatial, 948
  spatio-temporal, 958
  temporal, 1022, 948
  undetached, 1022, 948
  unextended, 948
  vs. component, 948
Part-of
  definability of, 27
  relation, 948
Part-whole relation, 947, 975
Parthood relation, 170
Parthood, 948
  and existence, 969
  and identity, 970
  and location theory, 951
  and vagueness, 971
  and 'part', 948
  Antisymmetry (P.3), 952
  antisymmetry of, 950, 952
  as a partial ordering, 950
  Atomicity (P.8), 962
  Atomlessness (P.7), 962
  basic principles, 950
  borderline cases of, 972
  Boundedness (P.10$_\psi$), 965
  Complementation (P.6), 960
  composition principles, 954
  conceptually prior to identity, 954
  connected, 983
  decomposition principles, 954
  definability of, 953, 968, 983
  Density (P.9), 964
  Determinacy (P.16), 975
  determinacy of, 975
  extensional, 1020, 1022, 957–959, 962, 967–968, 972, 985
  functional, 951
  Fusion (P.11$_\psi$), 965
  Fusion$_a$ (P.11$_{\psi a}$), 965
  Fusion$_b$ (P.11$_{\psi b}$), 965
  fuzzy, 972
  improper, 950
  indeterminacy of, 971
  Indiscernibility (P.3′), 954
  Infinitary Boundedness (P.12$_\xi$), 966
  Infinitary Fusion (P.13$_\xi$), 966
  isomorphic to set inclusion, 968
  proper, 950, 952, 955
  Reflexivity (P.1), 952
  reflexivity of, 950, 952

Restricted Atomicity (P.8$_\phi$), 963
Restricted Atomlessness (P.7$_\phi$), 963
Strong Atomicity (P.15), 971
Strong Company (P.4$_b$), 955
Strong Supplementation (P.5), 956
Supplementation (P.4), 955
time-indexed, 958
Transitivity (P.2), 952
transitivity of, 950, 952
Unique Unrestricted Fusion (P.14), 967
Unrestricted Fusion (P.13), 967
varieties of, 948
vs. conceptual inclusion, 948
vs. enclosure, 982, 999
vs. material constitution, 948, 951, 985
vs. mixture composition, 948
vs. part-whole relation, 947, 975, 979
Weak Company (P.4$_a$), 955
Partial order, 876
Partition, 39
connected, 39
graph, 48
manual, 731
radial, 45
c^h-, 42
Pasch, 345
axiom, 229, 388–389, 391, 397
Pasting, 476, 482
Path-connectedness, 988
Path-consistency, 165
algorithm, 166
method, 165
operation, 165
refinements, 186
Path-consistent, 165
Peirce's puzzle, 993
Pencil, 355
Penrose diagram, 697–698, 702
Percentiles, 192
Perception, 101
Perpendicularity, 349, 355
Persistent, 310
Phase portrait, 547
Phase-transition, 181
Phenomenology, 101
Photon, 613, 618, 621, 638, 651, 662, 668, 671, 677
Pieri Mario, 345, 395
relation, 394
PL-complex, 76
Place, 1013, 1025
absolute vs. relative, 1026
Plane graph
geometrical dual, 46
piecewise linear, 46
rational piecewise linear, 46
semi-algebraic, 46
Plane, 102, 360
Plotkin
powerdomain, 490
Poincaré Recurrence Theorem, 567, 573
Point structure, 821, 837
Point, 101–103, 106, 467
generalized, 439, 469
generic, 439
global, 440, 467
regular, 821
singular, 821
Point-based representation, 101–102
Pointless topology, 116
Points, 170, 348, 358
Polarity, 885
constraints, 182
Polygon, 20
Polygonal regions, 102
Polyhedral ball
definability of, 74–75
Polyhedron, 20
Polynomial, 805
linear, 813
multivariate, 813
Polytope, 20, 781
basic, 20
Pore, 864
Poset
directed complete, 439
Positional information, 177
Positive
direction, 204
logic, 436
Possibility operator, 108, 142, 509
Possible world, 143, 400
Post's correspondence problem, 552
Power lattice, 877
Powerdomain, 490
Powerlocale, 490
Powerobject, 465
Powerset
finite, 454
Practical efficiency, 190
Preconvex relations, 174
Predicate, 445
Predicative, 491
Preorder, 305
Preposition
spatial, 1025
Presheaf, 476, 480
Pretopos
∞-, 460–461
Prime ideal theorem, 84
Prime model, 86
Primitive, 101, 119, 135
Primitive recursive time, 556
Principle of equivalence, 663–664
Principle of relativity, 609, 661, 663

Product, 135
  frame, 236
  graphs, 718
  mereological, 967–968
  of lines, 358
  of modal logics, 558
Projective
  3D-space, 363
  extension, 366
  geometry, 345, 765
  plane, 1, 363
  space, 362
  validity, 363
Proper, 241
Propositional, 445
  encoding, 142
  satisfiability problem, 178
Proximity
  connection algebra, 132
    standard, 133
  relation, 118
  space, 117–118, 127, 129–130, 132, 151
    separated, 118, 133
  standard, 118
Pseudo-complement, 15, 33, 738
Pseudo-pullback, 469
Pythagorean field, 384
Q-cell, 76
  face vertex of, 79
  partition, 79
QEPCAD, 826
Qua object, 956
Quadrangle, 385
Qualitative Spatial Reasoning, 7, 119, 121, 134, 512
Qualitative
  distance, 176
  size relations, 199
Quantale, 753, 758
Quantification, 150
Quantifier elimination, 806
  by cylindrical algebraic decomposition, 826
  first-order logic with linear constraints, 824
  Fourier-Motzkin algorithm, 824
  Tarski's algorithm, 825
  first-order logic with polynomial constraints, 825
Quantifier-free, 103
Quantity, 618, 638
Quantum
  gravity, 613
  logic, 735
  mechanics, 613
Quasi-Boolean structure, 135
Quasi-equation, 752
Quasi-equational, 450
Quasi-parallel, 361
Quasi-tree, 248
Query language, 806–808, 811
  closed, 809
  closure property, 823
  computationally complete, 849, 851
  expressiveness, 829
  first-order, 800, 804
Query, 807
  composition, 823
  consistent, 808
  constraint, 808
  emptiness test, 823
  expressibility, 809
  membership test, 823
  plug-in evaluation, 822
  topological, 809, 832, 837, 851–852
  unrestricted, 808
Quotient
  simple geometric, 484
Randomly generated instances, 180, 191
Rank of space, 363
Rasiowa-Sikorski system, 147
Raster model, 810
Real field, 384
Real line, 442
  localic, 443
Real numbers, 610, 618–619, 629
Real root counting, 826
Real-closed field, 351, 384
Receptacle, 997
Rectangle algebra, 175
Reduct, 449, 460, 469, 471, 649
Reduction, 179
  by refinement, 186
  polynomial, 179
Reference frame, 618
  inertial, 609
Reference system, 609
Refinement, 166, 186
  basic matrix, 187
  matrix, 187
  strategy, 189
Reflection
  localic, 487
Reflects Alexandroff extensions, 252
Reflexive closure, 258
Region Connection Calculus, 69, 513, 1001, 1004, 1008, 102, 120–122, 124, 134, 139, 141, 146, 149, 170, 715, 737, 983, 999
  algebra, 124, 130
  RCC-8 relations, 121–122, 139, 141, 146, 170, 512
Region, 3, 512, 1012, 101–102, 106, 729, 752, 945, 952, 956–957
  as a set of points, 968
  extended, 956
  extensional, 957
  spatial, 1012, 945, 952, 956–957, 968
  unoccupied, 1016

vs. concrete spatial entity, 1012, 1019, 1021, 945, 957, 978, 986
geographic, 136
Region-based, 102
representation, 101–103, 140
Regionhood, 1016
Regular
category, 461
closed, 115, 119, 125, 130, 136, 140
algebra, 115, 129
set, 512
closure, 730
set, 512
open, 14, 115, 119, 125, 140, 727
algebra, 115
set, 115
Relation algebra, 103–104, 110, 112, 121, 149, 164, 654–655
operations on relations, 164
Relation, 103, 799, 807
attribute, 799
constraint, 807
database, 810
equivalent, 807
Euclidean, 331
finite, 804
finitely representable, 804
linear, 810
polynomial constraint, 807
semantics of a, 807
serial, 331
Relational
composition, 110
converse, 110
semantics, 305
Relationism, 1012, 1026, 945
Relative interior, 750
Relative, 623, 638
distance, 176
Relativity theory, 9, 607, 629, 641, 649–650, 657–659
general, 9, 607
special, 9, 607
general, 618, 620, 646–647, 657, 660–664, 676–677, 680, 683, 704, 706
special, 619, 630, 646–647, 657–658, 661–665, 670, 680, 705–706
Remainder
mereological, 1007, 1022, 955–956
Representation theorem, 109, 127, 129–130, 149
Residuation law, 287
Resolvable, 253
Respects components, 18, 20, 28
Restriction, 476
Retract, 719
Reverse mathematics, 608, 645, 659
Reverse operator, 206
Reverse relativity theory, 608, 645, 659
Rich language, 394
Riemann-tensor, 678
Ring, 350
Root, 241, 257
Rooted, 235, 241, 257
Rotation, 380
Rough sets, 898
adjunction, 900
duality, 901
fuzzy, 912
S-true, 254
S-valid, 254
S4◯, 568
S4C, 567
canonical dynamic Kripke model, 593
completeness for Kripke models, 593
completeness in Cantor Space, 569, 592, 599
completeness, 592, 594
finite model property, 592, 594
incompleteness in the Reals, 569, 592
the bimodal dynamic topological logic of continuous functions, 592
Sahlqvist theorem, 402
SAT, 178
Satisfaction, 805
Satisfiability, 178, 505
relation, 401
Satisfy, 459
Scattered, 253
Scenario, 167
Schema
database, 799, 807, 810–811
Schönflies theorem, 35, 71
polyhedral, 71
Scott
continuity, 464
Search space
size, 168
Second countable
topological space, 89
Section
local, 476
Selective unravelling, 242
Selective, 242
Self-connected, 151
Self-connectedness, 986, 988
Semantical, 649
consequence, 623, 647
Semantics
non-terminating, 843
of a linear tuple, 810
of a program, 850
of a relation, 807
of a transitive closure formula, 843, 848
terminating, 843
Semi-algebraic function, 815–816

Semi-algebraic set, 19, 803, 806–807, 812–816, 818–819, 822, 837, 846, 852
  approximation of a, 831, 847
  decomposition, 815
  linear approximation, 831
  plane graph, 51
Semi-linear set, 803, 806, 810, 812–813, 818, 841, 845, 849
Semi-regularisation, 115
Semilattice, 454
Sense data, 101
Sentence, 805
Separable, 233
Separated
  lightlike, 653
  spacelike, 653
  timelike, 653
Separating set, 461
Separation condition, 113, 130
  T_1, 114, 129, 134
  T_2, 114
  T_3, 115
  T_4, 115
  T_i-, 27
Separation number, 1006
Sequent, 436, 447
Serial subsets of $\mathbb{R}$
  logic of, 254
Serial, 227, 254, 353
Set of constraints, 164
Set theory
  fuzzy, 972
  non-well-founded, 951
Shadow, 1020
Sheaf, 462, 468, 474, 476–477, 482
Sheavification, 483
Shellable, 45
Shorter-than relation, 394
Sieve, 479
Signature, 433, 445
Similarly situated, 38
Simple functional extension, 484
Simple geometric quotient, 484
Simple, 241
  extended (mereological), 1021
  mereological, 961
Simplex, 822
Simultaneity, 613, 625–626, 642
Simultaneous, 613, 622, 625–626, 629, 632
Single-aspect calculi, 197
Singularity, 687, 691–692, 696–698, 701, 703
Site, 474, 479, 482
Skeleton, 248
Skew
  field, 351
  lines, 349, 381
Skin (topological), 1011
Slice, 478
Slope of line, 369
Slope, 611, 621, 625, 643, 668, 684, 692, 697, 699
Smooth, 662, 666–668, 674–678, 681, 700
Sober space, 442, 487
Social geometry, 10
Sort, 445, 618, 649, 651
Space
  as a fiction, 945
  atomicity of, 1017
  atomlessness of, 1017
  Bolzanian, 996
  boundarylessness of, 1017
  cumulativity of, 1017
  density of, 1017
  dimensionality of, 1006
  dissectiveness of, 1017
  Euclidean, 508, 1017, 991
  fullness of, 1016, 1021
  Leibnizian conception of, 1012, 945
  metric, 506
  modalities, 403
  Newtonian conception of, 1012, 945, 996
  perceptually remote, 945
  relationist conception of, 1012, 1026, 945
  substantivalist conception of, 1012, 1026, 945, 996
  Whiteheadian conception of, 956, 963, 984, 997
Space-dilation, 641
Space-isometry, 640, 646, 675
Space-time location, 624
Space-time, 9, 608, 617, 638, 642, 665
  location, 619–620, 624
  anti de Sitter, 703
  de Sitter, 703, 705
  Eddington-Finkelstein, 695–696
  general relativistic, 660–667, 670–672, 674, 676, 686
  Gödel's rotating universe, 701, 703, 705
  Kerr, 700
  Kruskal-Szekeres, 697
  Minkowski, 683–685, 688, 705
  Schwarzschild, 686–687, 694, 698
  Tipler-Stockum, 701
  vacuum, 705
Spacelike separated, 653
Span of lines, 357
Spatial entities
  points, 162, 173, 502
  regions, 162, 502
Spatial reasoning
  object-oriented, 946, 957, 986
  space-oriented, 946, 957, 986
  fusion, 917
  mathematical morphology, 903
  model-based recognition, 917
  qualitative, 512, 931

spatial relations, 903
Spatial regions, 170, 512
dimension, 171
holes, 171
internal connectedness, 171
non-emptiness, 171
regularity, 171
Spatial, 442
distance, 621
preposition, 1025
Spatio-temporal, 101, 150, 502
logic, 497
Special composition question, 965, 969
Special Principle of Relativity, 663–664
Specialization, 724
morphism, 463, 487
order, 238, 439, 472
Speed of light, 609, 621, 642, 663, 688, 696
Speed, 621, 677
Sphere, 32
Spherical geometry, 757, 770
Spherical oriented matroid, 771
Split set, 168
Splittable
mereotopology, 58
Splitting of objects, 991
Stage of definition, 439, 459
Stage, 467
Stalk, 474, 477
Standard
contact, 130
relation, 125
model, 434, 437, 440
product topology, 263
L_Σ-theory, 58, 69, 90
Star calculus, 174
Stereographic projection, 33
Stone
space, 435
theorem, 106, 127–128
Straight line, 612, 620, 629, 641, 643, 655, 662, 668, 680, 684, 692, 697, 699
String-theory, 618
Structuring element, 861
accessibility relation, 924
Structuring function, 873
Sturm theorem, 826
Subalgebra, 105
Sublocale, 491
Subobject classifier, 465
Subset frame, 308
Subspace, 233
Substantivalism, 1012, 1026, 945, 996
Substitution, 458
rule, 447
Sum, 718
axiom, 124
connected (mereological), 976
function, 135
mereological, 954, 968
Supervaluationism, 974
Supplement, 740
Supplementation
and complementation (mereological), 960
mereological, 955, 960, 967, 998
strong (mereological), 956, 966
Surface, 69, 1001, 747, 956, 977, 992
Surrounding set, 758
Swelling, 248
Symmetric difference, 110
Synchronism, 613, 616, 632, 637, 640, 642
Syntactic category, 451, 483
Syntactical, 649–650
Tangent line, 840
Tarski-Seidenberg theorem, 814
Tarski-Vaught lemma, 57
Temporal logic, 527
branching, 528
first-order, 557
linear, 527
Temporal modalities, 527, 566
henceforth, 527, 566
next, 527, 566
Temporal order, 614, 630
TERA project, 828
Term, 805, 811
in context, 446
Ternary
operation, 369
relation, 516
ring, 370
attached to plane, 370
linear, 373
Test instances, 191
Test space, 730
Theory, 351
ω-categorical, 352
Cartesian, 449
coherent type, 454
coherent, 436, 446
complete, 351
decidable, 352
essentially algebraic, 450
finitary algebraic, 449
geometric type, 453, 461
geometric, 436, 446, 484
left exact, 450
morphism, 446
quasi-equational, 450
Thought-experiment, 610, 615, 621, 627, 634, 637
Threshold set, 868
Tidal force, 692, 695
Time modalities, 403
Time travel, 661, 701

Time
  branching, 528
  inner, 670
  linear, 527
  proper, 656, 670
  wristwatch, 668, 670, 678, 689–691
Time-axis, 611, 620, 640–641, 673, 687, 690–691
Time-dilation, 632, 634–635, 637, 640, 690
Time-orientation, 647, 656, 700
Timelike, 668
  separated, 653
  curve, 668
Timewarp, 607
Tips, 412
Tolerance space, 719
Topo-bisimulation, 225
Topo-definable, 251
Topo-model, 218
Topologic, 314, 509
Topological space, 1, 106, 113, 127, 149, 231, 507, 565
  completely regular, 114, 134
  Hausdorff, 27, 114
  locally compact, 32
  locally connected, 18
  normal, 27, 114
  regular, 27, 114
  semi-regular, 18, 27, 114
  weakly normal, 114, 134
  weakly regular, 27, 114, 129–130
  connected, 134, 515
Topological
  bisimulation, 225
  definability, 251
  dynamics, 550, 566
  interior, 507, 565
  model, 218, 302, 510, 569
  product logic, 268
  property, 852
  query, 832, 837, 851
  semantics, 218, 565, 569
  sum, 233
  temporal model, 533
  type, 840
Topology, 301, 507, 664, 946
  Alexandroff, 306, 508
  and common sense, 991
  and mereology, 975
  and spatial reasoning, 946, 975
  intrinsic, 1010
  limits of, 1010
Topos theory
  fundamental theorem, 478
Topos, 9, 445
  elementary, 465
  Grothendieck, 460
  localic, 471, 487
Touching (topological), 981
Tractability, 150, 178
Transformation, 639, 643
  cut, 833–836
  glue, 833–836
  coordinate, 639–640
    Galilean, 640
    Newtonian, 640
  Lorentz, 637, 639–640, 675
  worldview, 639–640, 662
Transitive closure logic, 841–842, 845
  with start point and parameters, 848
  with stop conditions, 847
Transitive closure, 842
Translation, 260, 379
Trapezium relation, 375
Tree metric space, 522
Triangle relation, 349
Triangulation, 76, 761, 821
  theorem, 821, 847
Trivial, 232, 818
  semi-algebraically, 818
Triviality, 818
  theorem, 818–819, 840, 852
Trivialization, 818
Trivially flawed instances, 191
True, 234, 237
Truth in Kripke models, 401
Tuple, 799
  linear, 810
Turing
  barrier, 703
  computable, 849
  machine, 499
  reduction, 169
Twin paradox, 659–660, 670
Two-sorted
  Kripke models, 417
  modal language, 416
  modal logic for affine geometries, 420
  modal logic for projective geometries, 418
Type theory
  coherent, 454
  geometric, 452–453, 461, 463, 484
Ultra-contact, 126, 131
Ultrafilter, 84, 105, 126
Umbra, 868
Undecidability, 171
Under-constrained instances, 181
Underfilter, 891
Underlap
  mereological, 952
Uniform finiteness property, 817, 831
Uniformly finite, 817
Unit of measurement, 621–622, 653, 705
Unit
  line, 368

point, 368
Unit-vector, 627, 639, 646, 667, 673–674, 694–695
Universal entity, 952
Universal
quantification
finitely bounded, 454
region, 135
relation, 164, 166
Universalism
mereological, 971
Universe
mereological, 968
Upper powerlocale, 491
Upset, 238
V-open, 263
Vacuum, 701–703, 705
Vagueness, 742
of composition, 969
of parthood, 971
semantic vs. ontological, 973
de re vs. *de dicto*, 973
Valid, 234, 237
topologically, 303
Validity in Kripke frames, 401
Valuation, 234, 401
Variety, 107
graphs, 722
Vector
model, 810
space, 8, 10, 290, 620
Vertical topology, 263
Vertically open, 263
Vietoris powerlocale, 490
Visited nodes, 192
Visual field, 101
Vocabulary, 618, 648–650, 658–659, 676, 678, 805
Weak closure system, 726
Weakly transitive, 257
Well-powered, 457, 460
Wheel, 720
White hole, 697–698
Whitney's theorem, 49
Whole, 947, 975
scattered, 961, 969, 975, 987
Why-type question, 659
Windowing function, 888
Worldcurve, 678, 690
Worldline, 611, 620, 646, 662, 668, 670, 677–678, 680, 682, 690
Worldview relation, 618, 662
Worldview transformation, 639–641, 643, 645, 662, 680, 684
Worldview, 638–639, 643, 646, 648, 661–663, 677, 682, 684, 691
Wormhole, 607, 660–661, 676, 697, 704
Yoneda embedding, 480
2SAT, 179
3SAT, 179
4-intersection model, 136–137
9-intersection model, 137, 515
9-intersection, 172, 515