Martin Aigner
Günter M. Ziegler

Proofs from THE BOOK

Martin Aigner
Günter M. Ziegler

Proofs from THE BOOK

Edizione italiana
a cura di Alfio Quarteroni

Con 250 figure
Comprese le illustrazioni
di Karl H. Hofmann

Springer

MARTIN AIGNER
Freie Universität Berlin
Institut für Mathematik II (WE2)
Arnimallee 3
14195 Berlin, Germany
email: aigner@math.fu-berlin.de

GÜNTER M. ZIEGLER
Technische Universität Berlin
Institut für Mathematik, MA 6-2
Straße des 17. Juni 136
10623 Berlin, Germany
email: ziegler@math.tu-berlin.de

Edizione italiana a cura di:
ALFIO QUARTERONI
MOX, Politecnico di Milano, Milano, Italy
e CMCS-EPFL, Lausanne, Switzerland

Traduzione a cura di:
Silvia Quarteroni
Department of Computer Science
University of York, York, UK

Springer-Verlag fa parte di Springer Science+Business Media

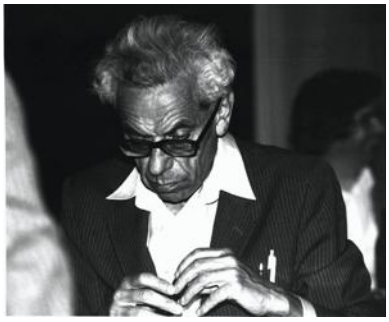
springer.com

ISBN 10 88-470-0435-7
ISBN 13 978-88-470-0435-7

Riprodotto in copia camera-ready da file originali forniti dagli Autori e rielaborati in italiano dal Traduttore
Progetto grafico della copertina: Deblik, Berlino
Stampato in Italia: Signum, Bollate (Mi)

Prefazione

Paul Erdős amava parlare del "Libro" in cui Dio conserva le dimostrazioni perfette per i teoremi matematici, seguendo il detto di G. H. Hardy secondo il quale non vi è posto perenne per la matematica brutta. Erdős diceva anche che non è necessario credere in Dio, tuttavia in quanto matematici si deve credere nel Libro. Alcuni anni fa gli suggerimmo di scrivere una prima (e assai modesta) approssimazione del Libro. Egli fu entusiasta dell'idea e, come gli era peculiare, si mise immediatamente al lavoro, riempiendo pagine su pagine con i suoi suggerimenti. Il nostro libro sarebbe dovuto essere pubblicato nel marzo 1998, come strenna per l'85-esimo compleanno di Erdős. Essendo sfortunatamente morto nell'estate del 1996, Paul non compare come co-autore. Tuttavia questo libro è dedicato alla sua memoria.



Paul Erdős

"Il Libro"

Non abbiamo alcuna definizione o caratterizzazione di cosa costituisca una *dimostrazione da Libro* (NdT: spesso lasceremo questa espressione nella sua versione originale inglese: *Proof from THE BOOK,* al fine di essere più fedeli al titolo di quest'opera): ci limiteremo qui a proporre alcuni esempi, nella speranza che i nostri lettori condividano il nostro entusiasmo per idee brillanti, astute intuizioni e meravigliose osservazioni. Speriamo che essi gradiscano tutto questo nonostante le imperfezioni della nostra esposizione. La scelta che abbiamo fatto è in larga misura influenzata dallo stesso Paul Erdős. Un buon numero di questi argomenti furono suggeriti direttamente da lui e molte delle dimostrazioni sono riconducibili direttamente a lui o furono abbozzate grazie al suo intuito supremo nel porre la giusta domanda o nel formulare la giusta congettura. Pertanto, questo libro riflette in larga misura il punto di vista di Paul Erdős su cosa debba considerarsi una *Proof from THE BOOK*.

Nella nostra scelta degli argomenti, la limitazione che ci siamo imposti è che qualunque cosa contenuta nel libro sia accessibile a lettori la cui formazione includa solo una modesta quantità di tecniche che si acquisiscono nei corsi di laurea in Matematica. Un po' di algebra lineare, alcuni elementi di base di analisi e teoria dei numeri, ed una salutare cucchiaiata di concetti e ragionamenti elementari di matematica discreta dovrebbero essere sufficienti per capire ed apprezzare tutto quanto vi è in questo libro.

Siamo estremamente riconoscenti alle tante persone che ci hanno aiutato e sostenuto in questo progetto — tra essi gli studenti di un seminario in cui abbiamo discusso una versione preliminare, Benno Artmann, Stephan Brandt, Stefan Felsner, Eli Goodman, Torsten Heldmann, e Hans Mielke. Ringraziamo Margrit Barrett, Christian Bressler, Ewgenij Gawrilow, Michael Joswig, Elke Pose, e Jörg Rambau per il loro aiuto tecnico nella fase

di composizione di questo testo. Abbiamo un grande debito nei confronti di Tom Trotter che ha letto il manoscritto dalla prima all'ultima pagina, di Karl H. Hofmann per i suoi magnifici disegni, e più di tutti nei confronti del grande Paul Erdős.

Berlino, marzo 1998 *Martin Aigner · Günter M. Ziegler*

Prefazione alla Seconda Edizione

La prima edizione di questo libro è stata accolta in modo meraviglioso. Inoltre, abbiamo ricevuto un numero inusuale di lettere contenenti commenti e correzioni, alcune scorciatoie, così come suggerimenti interessanti per dimostrazioni alternative e nuovi argomenti da trattare (pur cercando di riportare dimostrazioni *perfette*, tale non è la nostra esposizione).

La seconda edizione ci fornisce l'opportunità di presentare questa nuova versione del nostro libro: esso contiene tre ulteriori capitoli, revisioni sostanziali e nuove dimostrazioni in diversi altri capitoli, molte delle quali basate sui numerosi suggerimenti che abbiamo ricevuto. Abbiamo anche eliminato un capitolo del vecchio libro, quello sul "problema delle tredici sfere", la cui dimostrazione richiedeva dettagli che non abbiamo potuto completare in modo da renderla breve ed elegante.

Ringraziamo tutti i lettori che ci hanno scritto e pertanto ci hanno aiutato—fra essi Stephan Brandt, Christian Elsholtz, Jürgen Elstrodt, Daniel Grieser, Roger Heath-Brown, Lee L. Keener, Christian Lebœuf, Hanfried Lenz, Nicolas Puech, John Scholes, Bernulf Weißbach, e *molti* altri. Grazie di nuovo per l'aiuto e il supporto a Ruth Allewelt e Karl-Friedrich Koch di Springer Heidelberg, a Christoph Eyrich e Torsten Heldmann a Berlino, ed a Karl H. Hofmann per i nuovi splendidi disegni.

Berlino, settembre 2000 *Martin Aigner · Günter M. Ziegler*

Prefazione alla Terza Edizione

Non avremmo mai sognato, mentre preparavamo la prima edizione di questo libro nel 1998, il grande successo che questo progetto avrebbe avuto, con traduzioni in molte lingue, risposte entusiastiche da tanti lettori, e tanti meravigliosi consigli per miglioramenti, aggiunte, e nuovi argomenti — che potrebbero impegnarci per anni.

Dunque, questa terza edizione offre due nuovi capitoli (sulle identità delle partizioni di Eulero, e sul mescolamento delle carte), tre dimostrazioni sulla serie di Eulero appaiono in un capitolo a parte, e vi è un certo numero di ulteriori miglioramenti come il trattamento di Calkin-Wilf-Newman sulla "enumerazione dei razionali". Questo è tutto, per il momento!

Ringraziamo tutti coloro che hanno sostenuto questo progetto durante gli

ultimi cinque anni e il cui contributo ha reso questa nuova edizione diversa dalle precedenti. In particolare, David Bevan, Anders Björner, Dietrich Braess, John Cosgrave, Hubert Kalf, Günter Pickert, Alistair Sinclair, e Herb Wilf.

Berlino, luglio 2003 *Martin Aigner · Günter M. Ziegler*

Prefazione all'Edizione Italiana

Proofs from THE BOOK è un'opera straordinaria che ha saputo calamitare l'interesse di numerosissimi lettori, matematici e non, come poche altre di argomento matematico apparse in questi ultimi anni. Dall'edizione originale in lingua inglese, pubblicata nel 1998, sono poi state prodotte due altre edizioni in inglese e un numero in continua crescita di traduzioni in altre lingue. L'edizione italiana corrisponde alla traduzione della terza edizione del testo inglese, uscita nel 2004, fatte salve alcune piccole revisioni che ci sono state segnalate dagli stessi autori.

Proofs from THE BOOK rappresenta un'opera unica nel suo genere. La matematica è una disciplina costruita su teorie codificate in lemmi e teoremi le cui dimostrazioni sono sempre rigorose, spesso avvincenti e creative, talvolta bellissime. È proprio la tensione dei matematici di ogni epoca, che li spinge a cercare dimostrazioni belle, ad aver ispirato gli autori, i quali, immaginano che vi sia UN LIBRO, cioè THE BOOK (forse addirittura di ispirazione divina), che contenga le dimostrazioni più belle della matematica, quelle che rasentano la perfezione. Al fine di essere il più rispettosi possibile del suo solenne significato evocativo, abbiamo voluto mantenere il titolo originale dell'opera anche nell'edizione italiana.

Il testo tocca diversi campi della matematica, quali la teoria dei numeri, la geometria, l'analisi, la combinatoria, la teoria dei grafi, la probabilità; per ognuno di essi vengono scelti dei risultati particolarmente significativi che hanno marcato in modo irreversibile l'evoluzione di una disciplina antichissima e tuttora straordinariamente viva. Vengono proposte le dimostrazioni più geniali, avvincenti e belle – talvolta più dimostrazioni di ogni teorema – che a buon diritto si suppone trovino posto in THE BOOK.

Pur essendo un libro che tratta argomenti non banali di matematica, questo volume non è per soli matematici. Le sue dimostrazioni non richiedono a priori un approfondito bagaglio di conoscenze (in teoria, anche un bravo studente che abbia alle spalle una laurea in discipline scientifiche dovrebbe poterle capire, apprezzare e, soprattutto, gustare). Lo stile espositivo dell'edizione originale è teso alla ricerca dell'essenzialità e del rigore, in atteggiamento quasi riverente verso l'energia dirompente che questi teoremi e queste dimostrazioni sembrano sprigionare. Nella traduzione italiana abbiamo voluto rispettare questa impostazione austera, anche se il desiderio di aderire fedelmente all'originale abbia talvolta imposto la rinuncia alle rotondità tipiche del periodare italiano.

Una caratteristica saliente di quest'opera è la perfetta simbiosi fra dimostra-

zioni di risultati classici, quelli dovuti ad alcuni fra i giganti dello sviluppo delle conoscenze umane di diversi secoli fa – quali ad esempio Euclide, Eulero, Fermat, Bernoulli, Hermite, Cauchy – risultati che hanno segnato nel secolo scorso l'evoluzione verso la matematica moderna – Hilbert, Cantor, Ramanujan, Shannon, Turan, lo stesso Erdős, etc.– e risultati che rappresentano oggi il terminale applicativo di queste straordinarie teorie matematiche che si sono costantemente migliorate e generalizzate nel corso dei secoli. La matematica è infatti un albero vivo percorso in ogni sua componente (sino alle più piccole e moderne ramificazioni) da una linfa che si alimenta con continuità grazie a radici millenarie. Non stupisce allora che in questo libro, le teorie e i teoremi "del passato" trovino applicazione in ambiti di assoluta quotidianità: come può il direttore di un museo disporre il minor numero di guardie con la certezza che ogni sala venga sorvegliata; come assicurarsi che i croupier al casinò (o gli amici al bar) mescolino un mazzo di carte con la certezza che dopo un ben preciso numero di mescolamenti possano considerarsi distribuite in modo casuale; come mettere a punto una strategia che consenta di trovare accoppiamenti stabili fra ragazzi e ragazze oppure uomini e donne appartenenti a due gruppi diversi; a quale numero minimo di colori si debba ricorrere per colorare una carta geografica; come codificare informazioni complesse – audio o video – affinché vengano trasmesse senza errori con i moderni sistemi di telecomunicazione; come completare un quadrato latino, una sorta di predecessore del moderno Sudoku; come spiegare razionalmente il fatto che il direttore del famoso hotel di Hilbert, quello con infinite stanze, possa trovar posto a nuovi ospiti anche quando l'albergo sia al completo; e innumerevoli altri.

Ritengo che questa costante interrelazione ed interposizione fra classico e moderno, teorico e applicato, finito ed infinito, nonché la presenza di numerose illustrazioni del geniale Karl H. Hofmann, non possano non affascinare ed intrigare anche chi, pur non essendo matematico, abbia sempre manifestato curiosità (o interesse) verso la più nobile e fondamentale delle scienze moderne.

Desidero, insieme a Martin Aigner e Günter M. Ziegler, ringraziare calorosamente Silvia Quarteroni del Computer Science Department dell'Università di York per la traduzione di quest'opera, seguita con passione e cura dei minimi particolari e Gianluigi Rozza dell'École Polytechnique Fédérale di Losanna per essersi occupato con generosità e precisione di tutti gli aspetti relativi alla gestione tecnica dei file ed alla correzione delle bozze. Desidero infine ringraziare i miei colleghi del Politecnico di Milano, i professori Alessandra Cherubini, Daniela Lupo, Marco Fuhrman e Piercesare Secchi, che mi hanno aiutato a trovare la traduzione più appropriata di alcuni termini matematici astrusi in alcuni settori di loro competenza.

Milano, novembre 2005 *Alfio Quarteroni*

Nota dell'Editore

Paul Erdős era un genio; personaggio istrionico, è rimasto senza lavoro per la maggior parte della sua vita, contando pertanto sull'ospitalità di istituzioni e di colleghi alla cui porta spesso bussava alle ore più improbabili dichiarando che "la sua mente era aperta". Pur avendo condotto una vita errabonda e stravagante, Erdős è considerato una delle menti matematiche più grandiose del XX secolo, che ha fatto della ricerca dell'eleganza la caratteristica preminente del suo lavoro.

Pubblicare un volume di questo tipo, ispirato alla filosofia di vita di Paul Erdős, è stato per Springer una sfida accattivante, poiché è notorio quanto sia difficile promuovere un certo tipo di matematica, cosiddetta "divulgativa". In una recensione di questo volume apparsa su Zentralblatt Math nel 2002, Juergen Appell asseriva che è opportuno ringraziare Springer per l'insolita decisione di pubblicare anche l'edizione tedesca del famoso libro, apparso inizialmente in lingua inglese e venduto con successo in tutto il mondo. Ebbene, dalla data di pubblicazione della prima edizione in inglese ad oggi, il LIBRO, che ha venduto globalmente circa 35.000 copie, è stato tradotto in svariate lingue, precisamente in tedesco, francese, giapponese, polacco, portoghese, ungherese, farsi, mentre sono di imminente pubblicazione le traduzioni in russo, spagnolo, coreano, turco e, probabilmente, anche in cinese.

Non poteva quindi mancare l'edizione italiana. In considerazione del forte significato scientifico, ma anche editoriale, del LIBRO, abbiamo deciso di affidare questa traduzione alle sapienti cure di uno dei nomi di maggiore spessore nell'ambito della matematica italiana ed internazionale, il Prof. Alfio Quarteroni; cogliamo qui l'occasione di ringraziarlo non solo per avere accettato di imbarcarsi in questa avventura impegnativa, ma anche per la precisione e la raffinatezza con cui ha seguito i minimi dettagli dell'operazione.

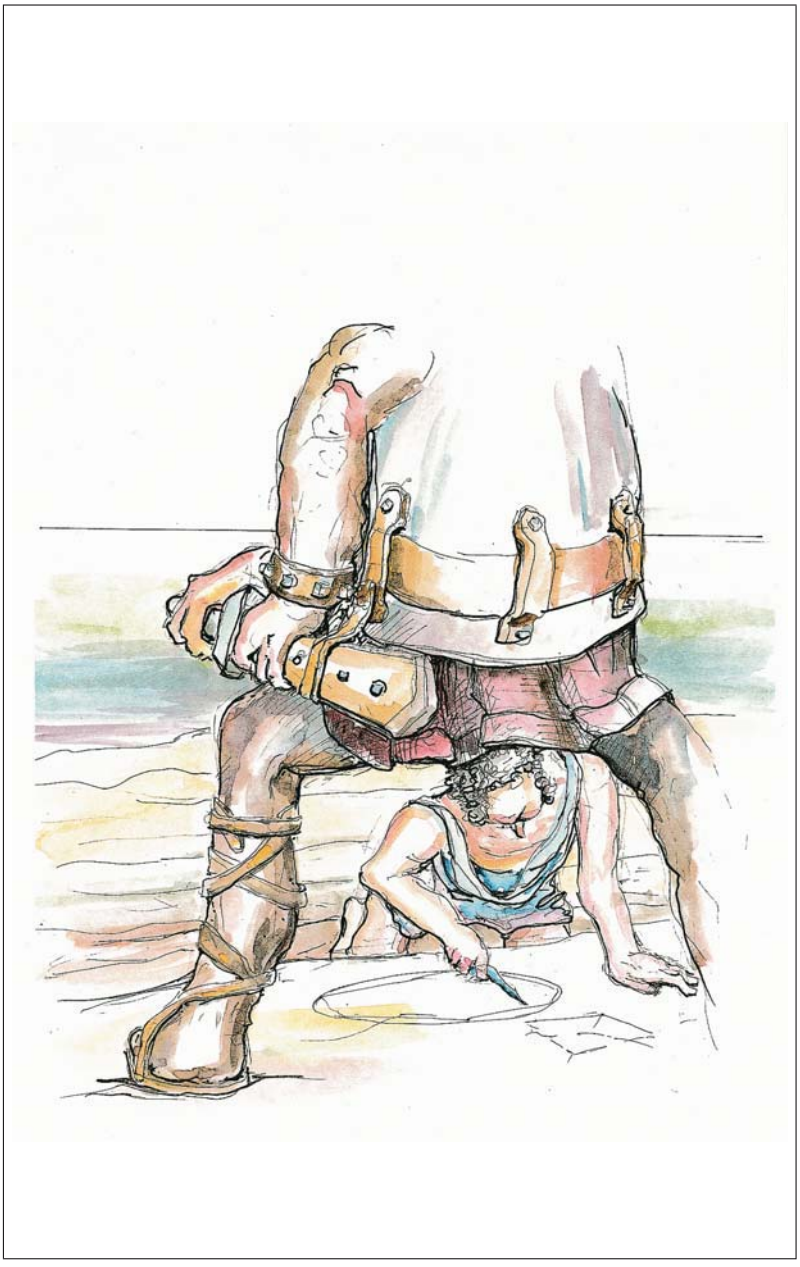
Abbiamo voluto mantenere l'inusuale formato dell'edizione originale inglese con gli ampi margini voluti per dare rilievo ad alcune definizioni ed esempi di particolare interesse, oltre che ai deliziosi schizzi di Karl H. Hofmann; riteniamo che questo libro non sia solo piacevole da leggere, ma anche da tenere in mano e guardare. Ci auguriamo che la nostra traduzione del LIBRO riscuota lo stesso successo avuto all'estero e che sia spunto per i matematici italiani per nuove discussioni su cosa sia bello, cosa sia arguto o cosa sia, come piacerebbe a Erdős, elegante.

Milano, novembre 2005

Francesca Bonadei
Springer-Verlag Italia

Sommario

Teoria dei Numeri



"Irrazionalità e π"

I numeri primi sono infiniti: sei dimostrazioni

Capitolo 1

È del tutto naturale incominciare queste note con quella che probabilmente è la più antica *Book Proof*, generalmente attribuita ad Euclide (*Elementi* IX, 20), la quale mostra che la successione dei numeri primi non è limitata.

■ **Dimostrazione di Euclide.** Per ogni insieme finito $\{p_1, \ldots, p_r\}$ di numeri primi, consideriamo il numero $n = p_1 p_2 \cdots p_r + 1$. Questo n ha un divisore primo p. Tuttavia p non è uno dei p_i: in caso contrario, p sarebbe un divisore di n e del prodotto $p_1 p_2 \cdots p_r$, dunque anche della differenza $n - p_1 p_2 \ldots p_r = 1$, il che è impossibile. Pertanto un insieme finito $\{p_1, \ldots, p_r\}$ non può rappresentare la collezione di *tutti* i numeri primi. □

Prima di continuare, introduciamo alcune notazioni. $\mathbb{N} = \{1, 2, 3, \ldots\}$ è l'insieme dei numeri naturali, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ l'insieme degli interi, $\mathbb{P} = \{2, 3, 5, 7, \ldots\}$ quello dei numeri primi.

Nel seguito, presenteremo varie altre dimostrazioni (tratte da una lista molto più lunga); speriamo piacciano al lettore tanto quanto piacciono a noi. Nonostante utilizzino diversi punti di vista, l'idea di base è comune a tutte: i numeri naturali crescono al di là di ogni possibile limite e ogni numero naturale $n \geq 2$ ha un divisore primo. Questi due fatti considerati insieme obbligano $\mathbb{P}$ ad essere infinito. La dimostrazione che segue è dovuta a Christian Goldbach (ed è tratta da una lettera a Eulero del 1730), la terza dimostrazione fa parte del folclore, la quarta è dovuta allo stesso Eulero, la quinta fu proposta da Harry Fürstenberg, mentre l'ultima è dovuta a Paul Erdős.

■ **Seconda dimostrazione.** Consideriamo dapprima i *numeri di Fermat* $F_n = 2^{2^n} + 1$ per $n = 0, 1, 2, \ldots$. Mostreremo che ogni coppia di numeri di Fermat è composta da numeri primi fra loro; ne seguirà che debbono esistere infiniti numeri primi. A questo scopo, verifichiamo la relazione ricorsiva

$$\prod_{k=0}^{n-1} F_k \;=\; F_n - 2 \qquad (n \geq 1),$$

dalla quale la nostra tesi segue immediatamente. In effetti, se m ad esempio è un divisore di F_k ed F_n $(k < n)$, allora m divide 2, perciò $m = 1$ o 2. Ma $m = 2$ è impossibile dal momento che tutti i numeri di Fermat sono dispari.

Dimostriamo la precedente relazione ricorsiva procedendo per induzione su n. Per $n = 1$ abbiamo $F_0 = 3$ e $F_1 - 2 = 3$. Per induzione concludiamo

$$\begin{aligned} F_0 &= 3 \\ F_1 &= 5 \\ F_2 &= 17 \\ F_3 &= 257 \\ F_4 &= 65537 \\ F_5 &= 641 \cdot 6700417 \end{aligned}$$

I primi sei numeri di Fermat

Teorema di Lagrange

Se G è un gruppo (moltiplicativo) finito e U è un sottogruppo, allora $|U|$ divide $|G|$.

■ **Dimostrazione.** Consideriamo la relazione binaria

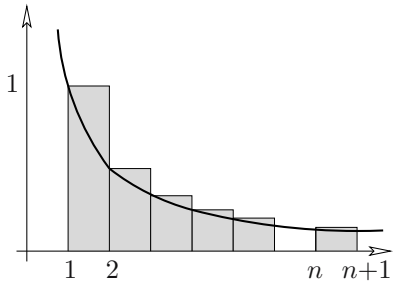
$$a \sim b : \iff ba^{-1} \in U.$$

Segue dagli assiomi di gruppo che $\sim$ è una relazione di equivalenza. La classe di equivalenza che contiene un elemento a è precisamente il coinsieme

$$Ua = \{xa : x \in U\}.$$

Poiché chiaramente $|Ua| = |U|$, troviamo che G si scompone in classi di equivalenza, tutte di grandezza $|U|$, e pertanto che $|U|$ divide $|G|$. □

Nel caso particolare in cui U sia un sottogruppo ciclico, $\{a, a^2, \ldots, a^m\}$ troviamo che m (il più piccolo intero positivo tale che $a^m = 1$, detto l'*ordine* di a) divide la cardinalità $|G|$ del gruppo.



Gradini sopra la funzione $f(t) = \frac{1}{t}$

che

$$\prod_{k=0}^{n} F_k \;=\; \Big(\prod_{k=0}^{n-1} F_k\Big)F_n \;=\; (F_n - 2)F_n \;=$$
$$= (2^{2^n} - 1)(2^{2^n} + 1) \;=\; 2^{2^{n+1}} - 1 \;=\; F_{n+1} - 2. \qquad \square$$

■ **Terza dimostrazione.** Supponiamo che $\mathbb{P}$ sia finito e p rappresenti il più grande numero primo. Consideriamo il cosiddetto *numero di Mersenne* $2^p - 1$ e mostriamo che ogni fattore primo q di $2^p - 1$ è maggiore di p, il che permetterà di ottenere la conclusione desiderata. Sia q un numero primo divisore di $2^p - 1$, pertanto $2^p \equiv 1 \pmod q$. Essendo p primo, ciò significa che l'elemento 2 ha ordine p nel gruppo moltiplicativo $\mathbb{Z}_q \backslash \{0\}$ del campo $\mathbb{Z}_q$. Questo gruppo ha $q - 1$ elementi. Dal teorema di Lagrange (riportato nel riquadro) sappiamo che l'ordine di ogni elemento divide la cardinalità del gruppo, ovvero abbiamo $p \mid q - 1$, pertanto $p < q$. □

Vediamo ora una dimostrazione che usa l'analisi elementare.

■ **Quarta dimostrazione.** Sia x un numero reale e $\pi(x) := \#\{p \le x : p \in \mathbb{P}\}$ il numero di numeri primi minori o uguali a x. Ordiniamo in modo crescente i numeri primi $\mathbb{P} = \{p_1, p_2, p_3, \ldots\}$. Consideriamo il logaritmo naturale $\log x$, definito come $\log x = \int_1^x \frac{1}{t} dt$.

Confrontiamo ora l'area soggiacente al grafico di $f(t) = \frac{1}{t}$ con una funzione a gradino maggiorante (si veda anche l'appendice a pagina 10 per questo metodo). Allora per $n \le x < n + 1$ abbiamo

$$\begin{aligned} \log x &\le 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1} + \frac{1}{n} \\ &\le \sum \frac{1}{m}, \text{ dove la somma si estende a tutti gli } m \in \mathbb{N} \text{ che} \\ &\qquad \text{hanno solo divisori primi } p \le x. \end{aligned}$$

Poiché ogni m di questo tipo può essere scritto in modo *univoco* come un prodotto della forma $\prod_{p \le x} p^{k_p}$, si vede che quest'ultima somma è uguale a

$$\prod_{\substack{p \in \mathbb{P} \\ p \le x}} \Big(\sum_{k \ge 0} \frac{1}{p^k}\Big).$$

La somma interna è una serie geometrica di ragione $\frac{1}{p}$, pertanto

$$\log x \;\le\; \prod_{\substack{p \in \mathbb{P} \\ p \le x}} \frac{1}{1 - \frac{1}{p}} \;=\; \prod_{\substack{p \in \mathbb{P} \\ p \le x}} \frac{p}{p-1} \;=\; \prod_{k=1}^{\pi(x)} \frac{p_k}{p_k - 1}.$$

Poiché, chiaramente, $p_k \ge k + 1$, otteniamo

$$\frac{p_k}{p_k - 1} \;=\; 1 + \frac{1}{p_k - 1} \;\le\; 1 + \frac{1}{k} \;=\; \frac{k+1}{k},$$

e pertanto

$$\log x \;\leq\; \prod_{k=1}^{\pi(x)} \frac{k+1}{k} \;=\; \pi(x)+1.$$

Tutti sanno che $\log x$ non è limitato, dunque $\pi(x)$ è anch'esso illimitato, così possiamo concludere che ci sono infiniti numeri primi. □

■ **Quinta dimostrazione.** Dopo l'analisi, è la volta della topologia! Consideriamo la seguente curiosa topologia sull'insieme $\mathbb{Z}$ degli interi. Per $a, b \in \mathbb{Z}$, $b > 0$, poniamo

$$N_{a,b} = \{a + nb : n \in \mathbb{Z}\}.$$

Ogni insieme $N_{a,b}$ è una progressione aritmetica infinita nei due sensi. Diciamo che un insieme $O \subseteq \mathbb{Z}$ è *aperto* se esso è vuoto, oppure se per ogni $a \in O$ esiste $b > 0$ con $N_{a,b} \subseteq O$. Naturalmente, l'unione di insiemi aperti è ancora un aperto. Se O_1, O_2 sono aperti, e $a \in O_1 \cap O_2$ con $N_{a,b_1} \subseteq O_1$ ed $N_{a,b_2} \subseteq O_2$, allora $a \in N_{a,b_1 b_2} \subseteq O_1 \cap O_2$. Dunque concludiamo che ogni intersezione finita di aperti è un aperto. Pertanto, questa famiglia di insiemi aperti induce una topologia in bona fide (NdT: in latino nella versione originale) su $\mathbb{Z}$.

Valgono le due seguenti proprietà:

(A) Ogni insieme aperto non vuoto è infinito.

(B) Ogni insieme $N_{a,b}$ è anche chiuso.

In effetti, la prima segue dalla definizione. Quanto alla seconda, osserviamo che

$$N_{a,b} \;=\; \mathbb{Z} \setminus \bigcup_{i=1}^{b-1} N_{a+i,b},$$

il che prova che $N_{a,b}$ è il complementare di un insieme aperto e pertanto è chiuso.

Sino ad ora i numeri non sono ancora entrati in scena — ma eccoli arrivare. Poiché ogni numero $n \neq 1, -1$ ha un divisore primo p, e dunque è contenuto in $N_{0,p}$, concludiamo che

$$\mathbb{Z} \setminus \{1, -1\} \;=\; \bigcup_{p \in \mathbb{P}} N_{0,p}.$$

Ora se $\mathbb{P}$ fosse finito, allora $\bigcup_{p\in\mathbb{P}} N_{0,p}$ sarebbe l'unione finita di insiemi chiusi (grazie a (B)), pertanto sarebbe chiuso. Conseguentemente, l'insieme $\{1, -1\}$ sarebbe aperto, ma ciò violerebbe (A). □

"Far rimbalzare sassolini all'infinito"

■ **Sesta dimostrazione.** La nostra dimostrazione finale si spinge considerevolmente oltre e dimostra non solo che esistono infiniti numeri primi, ma anche che la serie $\sum_{p\in\mathbb{P}} \frac{1}{p}$ diverge. Questo importante risultato fu formulato per la prima volta da Eulero (il che lo rende già di per sé interessante), ma la nostra dimostrazione, concepita da Erdős, è di una bellezza travolgente.

Sia $p_1, p_2, p_3, \ldots$ la successione dei numeri primi in ordine crescente, e supponiamo che $\sum_{p\in\mathbb{P}} \frac{1}{p}$ converga. Allora deve esistere un numero naturale k tale che $\sum_{i\geq k+1} \frac{1}{p_i} < \frac{1}{2}$. Chiamiamo $p_1, \ldots, p_k$ i numeri primi *piccoli*, e $p_{k+1}, p_{k+2}, \ldots$ i *grandi*. Per un arbitrario numero naturale N troviamo

$$\sum_{i\geq k+1} \frac{N}{p_i} \;<\; \frac{N}{2}. \tag{1}$$

Sia N_b il numero degli interi positivi $n \leq N$ che sono divisibili per almeno un numero primo grande, e N_s il numero di interi positivi $n \leq N$ che hanno soltanto divisori primi piccoli. Vogliamo mostrare che per un opportuno N

$$N_b + N_s \;<\; N,$$

il che rappresenterà la contraddizione cercata, in quanto per definizione $N_b + N_s$ dovrebbe essere uguale ad N.

Per stimare N_b notiamo che $\lfloor \frac{N}{p_i} \rfloor$ conta il numero di interi positivi $n \leq N$ che sono multipli di p_i. Pertanto dalla (1) troviamo

$$N_b \;\leq\; \sum_{i\geq k+1} \left\lfloor \frac{N}{p_i} \right\rfloor \;<\; \frac{N}{2}. \tag{2}$$

Consideriamo ora N_s. Scriviamo ogni $n \leq N$ che ha solo divisori primi piccoli nella forma $n = a_n b_n^2$, dove a_n è la parte non quadrata. Ogni a_n è perciò il prodotto di numeri primi piccoli *diversi* fra loro, e concludiamo che esistono precisamente 2^k parti non quadrate. Inoltre, poiché $b_n \leq \sqrt{n} \leq \sqrt{N}$, troviamo che ci sono al più $\sqrt{N}$ diverse parti quadrate, e pertanto

$$N_s \;\leq\; 2^k\sqrt{N}.$$

Poiché (2) vale per *ogni* N, resta da trovare un numero N per il quale $2^k\sqrt{N} \leq \frac{N}{2}$ oppure $2^{k+1} \leq \sqrt{N}$, e a tale scopo è sufficiente prendere $N = 2^{2k+2}$. □

Bibliografia

[1] B. ARTMANN: *Euclid — The Creation of Mathematics,* Springer-Verlag, New York 1999.

[2] P. ERDŐS: *Über die Reihe* $\sum \frac{1}{p}$, Mathematica, Zutphen B **7** (1938), 1-2.

[3] L. EULER: *Introductio in Analysin Infinitorum,* Tomus Primus, Lausanne 1748; Opera Omnia, Ser. 1, Vol. 8.

[4] H. FÜRSTENBERG: *On the infinitude of primes,* Amer. Math. Monthly **62** (1955), 353.

Il postulato di Bertrand

Capitolo 2

Abbiamo visto che la successione dei numeri primi $2, 3, 5, 7, \ldots$ è infinita. Per verificare che le distanze tra numeri primi successivi sono illimitate, si denoti con $N := 2 \cdot 3 \cdot 5 \cdot \ldots \cdot p$ il prodotto di tutti i numeri primi più piccoli di $k + 2$; si noti che nessuno dei k numeri

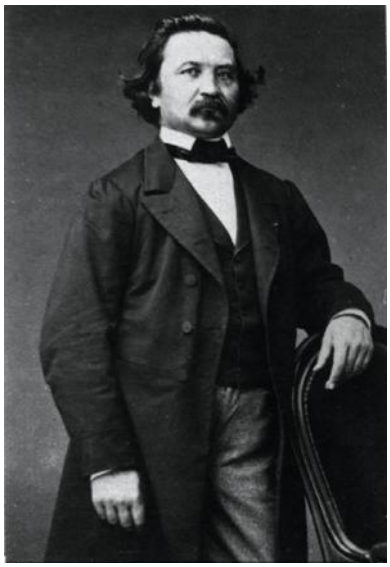
$$N + 2, N + 3, N + 4, \ldots, N + k, N + (k + 1)$$

è primo, poiché per $2 \leq i \leq k + 1$ sappiamo che i ha un fattore primo minore di $k + 2$ e tale fattore divide anche N, dunque anche $N + i$. Con questa ricetta troviamo, per esempio per $k = 10$, che nessuno dei dieci numeri

$$2312, 2313, 2314, \ldots, 2321$$

è primo.

Non solo, possiamo anche ottenere delle maggiorazioni per le suddette distanze nella successione dei numeri primi. Una famosa maggiorazione stabilisce che "La distanza fra un numero primo ed il suo successivo non può mai superare il più piccolo dei due". Tale enunciato è noto come postulato di Bertrand, poiché fu formulato e verificato empiricamente per $n < 3\,000\,000$ da Joseph Bertrand. Esso fu dimostrato per la prima volta per tutti gli n da Pafnuty Chebyshev nel 1850. Una dimostrazione molto più semplice fu data dal genio indiano Ramanujan. La nostra *Book Proof* è di Paul Erdős: essa è tratta dalla sua prima pubblicazione, apparsa nel 1932, quando Erdős aveva 19 anni.



Joseph Bertrand

Il postulato di Bertrand.
Per ogni $n \geq 1$, esiste un numero primo p tale che $n < p \leq 2n$.

■ **Dimostrazione.** Stimeremo la grandezza del coefficiente binomiale $\binom{2n}{n}$ con sufficiente precisione da verificare che se esso non avesse alcun fattore primo nell'intervallo $n < p \leq 2n$, allora sarebbe "troppo piccolo". Il nostro ragionamento si articola in cinque fasi.

(1) Dimostriamo dapprima il postulato di Bertrand per $n < 4000$. A questo scopo non è necessario verificare 4000 casi: è sufficiente (si tratta del "trucco di Landau") controllare che

$$2, 3, 5, 7, 13, 23, 43, 83, 163, 317, 631, 1259, 2503, 4001$$

Beweis eines Satzes von Tschebyschef.

Von P. Erdős in Budapest.

Für den zuerst von Tschebyschef bewiesenen Satz, laut dessen es zwischen einer natürlichen Zahl und ihrer zweifachen stets wenigstens eine Primzahl gibt, liegen in der Literatur mehrere Beweise vor. Als einfachsten kann man ohne Zweifel den Beweis von Ramanujan[1]) bezeichnen. In seinem Werk *Vorlesungen über Zahlentheorie* (Leipzig, 1927), Band I, S. 66—68 gibt Herr Landau einen besonders einfachen Beweis für einen Satz über die Anzahl der Primzahlen unter einer gegebenen Grenze, aus welchem unmittelbar folgt, daß für ein geeignetes q zwischen einer natürlichen Zahl und ihrer q-fachen stets eine Primzahl liegt. Für die augenblicklichen Zwecken des Herrn Landau kommt es nicht auf die numerische Bestimmung der im Beweis auftretenden Konstanten an; man überzeugt sich aber durch eine numerische Verfolgung des Beweises leicht, daß q jedenfalls größer als 2 ausfällt.

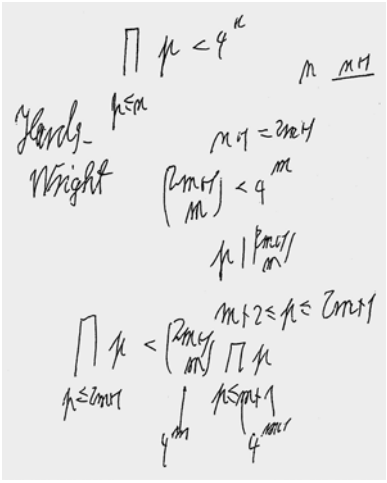
In den folgenden Zeilen werde ich zeigen, daß man durch eine Verschärfung der dem Landauschen Beweis zugrunde liegenden Ideen zu einem Beweis des oben erwähnten Tschebyschefschen Satzes gelangen kann, der — wie mir scheint — an Einfachkeit nicht hinter dem Ramanujanschen Beweis steht. Griechische Buchstaben sollen im Folgenden durchwegs positive, lateinische Buchstaben natürliche Zahlen bezeichnen; die Bezeichnung p ist für Primzahlen vorbehalten.

1. Der Binomialkoeffizient

$$\binom{2a}{a} = \frac{(2a)!}{(a!)^2}$$

[1]) Srr. Ramanujan, A Proof of Bertrand's Postulate, *Journal of the Indian Mathematical Society*, 11 (1919), S. 181—182 — *Collected Papers of* Srinivasa Ramanujan (Cambridge, 1927), S. 208—209.

sia una successione di numeri primi in cui ciascun numero è minore del doppio di quello che lo precede. Pertanto ogni intervallo $\{y : n < y \le 2n\}$, con $n \le 4000$, contiene uno di questi 14 numeri primi.



(2) Proviamo ora che

$$\prod_{p\le x} p \;\le\; 4^{x-1} \qquad \text{per ogni reale } x \ge 2, \tag{1}$$

dove la nostra notazione — qui e nel seguito — significa che il prodotto è esteso a tutti i numeri *primi* $p \le x$. La nostra dimostrazione procede per induzione sul numero di questi numeri primi. Essa non è tratta dalla pubblicazione originale di Erdős, tuttavia è anch'essa dovuta ad Erdős (si veda a margine) ed è una vera *Book Proof*. Innanzitutto osserviamo che se q è il più grande numero primo tale che $q \le x$, allora

$$\prod_{p\le x} p \;=\; \prod_{p\le q} p \qquad \text{e} \qquad 4^{q-1} \;\le\; 4^{x-1}.$$

Pertanto è sufficiente verificare (1) per il caso in cui $x = q$ sia un numero primo. Per $q = 2$ si ha "$2 \le 4$," dunque continuiamo considerando i numeri primi dispari nella forma $q = 2m + 1$. (Qui possiamo assumere, per induzione, che (1) sia valida per tutti gli interi x nell'insieme $\{2, 3, \ldots, 2m\}$.) Per $q = 2m + 1$ scomponiamo il prodotto e calcoliamo

$$\prod_{p\le 2m+1} p \;=\; \prod_{p\le m+1} p \cdot \prod_{m+1<p\le 2m+1} p \;\le\; 4^m \binom{2m+1}{m} \;\le\; 4^m 2^{2m} \;=\; 4^{2m}.$$

Si tratta di passaggi di facile verifica. In effetti,

$$\prod_{p\le m+1} p \;\le\; 4^m$$

si ottiene per induzione.

La disuguaglianza

$$\prod_{m+1<p\le 2m+1} p \;\le\; \binom{2m+1}{m}$$

segue dall'osservazione che $\binom{2m+1}{m} = \frac{(2m+1)!}{m!(m+1)!}$ è un intero, in cui i numeri primi che consideriamo sono tutti fattori del numeratore $(2m+1)!$, ma non del denominatore $m!(m+1)!$. Infine

$$\binom{2m+1}{m} \;\le\; 2^{2m}$$

vale in quanto

$$\binom{2m+1}{m} \text{ e } \binom{2m+1}{m+1}$$

sono due addendi (identici!) che compaiono in

$$\sum_{k=0}^{2m+1} \binom{2m+1}{k} = 2^{2m+1}.$$

(3)
Dal teorema di Legendre (si veda il riquadro) otteniamo che $\binom{2n}{n} = \frac{(2n)!}{n!n!}$ contiene il fattore primo p esattamente

$$\sum_{k\geq 1}\left(\left\lfloor\frac{2n}{p^k}\right\rfloor - 2\left\lfloor\frac{n}{p^k}\right\rfloor\right)$$

volte. Ogni addendo della somma vale al più 1, in quanto si ha

$$\left\lfloor\frac{2n}{p^k}\right\rfloor - 2\left\lfloor\frac{n}{p^k}\right\rfloor \;<\; \frac{2n}{p^k} - 2\left(\frac{n}{p^k}-1\right) \;=\; 2,$$

ed è un numero intero. Inoltre gli addendi si annullano quando $p^k > 2n$. Pertanto $\binom{2n}{n}$ contiene p esattamente

$$\sum_{k\geq 1}\left(\left\lfloor\frac{2n}{p^k}\right\rfloor - 2\left\lfloor\frac{n}{p^k}\right\rfloor\right) \;\leq\; \max\{r : p^r \leq 2n\}$$

volte. Ne segue che la più grande potenza di p che divide $\binom{2n}{n}$ non è più grande di $2n$. In particolare, i numeri primi $p > \sqrt{2n}$ compaiono al massimo una sola volta in $\binom{2n}{n}$.
Inoltre — e questo, secondo Erdős, rappresenta la chiave di volta della sua dimostrazione — i numeri primi p che soddisfano $\frac{2}{3}n < p \leq n$ non dividono affatto $\binom{2n}{n}$. In effetti, $3p > 2n$ implica (per $n \geq 3$, e dunque $p \geq 3$) che p e $2p$ sono i soli multipli di p che appaiono come fattori nel numeratore di $\frac{(2n)!}{n!n!}$, mentre abbiamo due p-fattori nel denominatore.

(4) Ora possiamo stimare $\binom{2n}{n}$. Per $n \geq 3$, usando una minorazione data a pagina 12, otteniamo

$$\frac{4^n}{2n} \;\leq\; \binom{2n}{n} \;\leq\; \prod_{p\leq\sqrt{2n}} 2n \;\cdot \prod_{\sqrt{2n}<p\leq\frac{2}{3}n} p \;\cdot \prod_{n<p\leq 2n} p$$

e pertanto, essendoci non più di $\sqrt{2n}$ numeri primi $p \leq \sqrt{2n}$,

$$4^n \;\leq\; (2n)^{1+\sqrt{2n}} \;\cdot \prod_{\sqrt{2n}<p\leq\frac{2}{3}n} p \;\cdot \prod_{n<p\leq 2n} p \qquad \text{per} \quad n \geq 3. \qquad (2)$$

(5) Supponiamo ora che non vi siano numeri primi p tali che $n < p \leq 2n$, così che il secondo prodotto in (2) valga 1. Sostituendo (1) in (2) otteniamo

$$4^n \;\leq\; (2n)^{1+\sqrt{2n}}\, 4^{\frac{2}{3}n}$$

ovvero

$$4^{\frac{1}{3}n} \;\leq\; (2n)^{1+\sqrt{2n}}, \qquad (3)$$

il che è falso per n sufficientemente grande! In effetti, usando la disuguaglianza $a+1 < 2^a$ (che si dimostra per induzione per tutti i numeri $a \geq 2$) si ottiene

$$2n = \left(\sqrt[6]{2n}\right)^6 < \left(\left\lfloor\sqrt[6]{2n}\right\rfloor + 1\right)^6 < 2^{6\left\lfloor\sqrt[6]{2n}\right\rfloor} \leq 2^{6\sqrt[6]{2n}}, \qquad (4)$$

Teorema di Legendre

Il numero $n!$ contiene il fattore primo p esattamente

$$\sum_{k\geq 1}\left\lfloor\frac{n}{p^k}\right\rfloor$$

volte.

■ **Dimostrazione.** Vi sono esattamente $\lfloor\frac{n}{p}\rfloor$ fattori di $n! = 1\cdot 2\cdot 3\cdot\ldots\cdot n$ divisibili per p, il che si traduce in $\lfloor\frac{n}{p}\rfloor$ p-fattori. Inoltre, $\lfloor\frac{n}{p^2}\rfloor$ fattori di $n!$ sono anche divisibili per p^2, il che si traduce nei successivi $\lfloor\frac{n}{p^2}\rfloor$ fattori primi p di $n!$, etc. □

I seguenti esempi
$\binom{26}{13} = 2^3\cdot 5^2\cdot 7\cdot 17\cdot 19\cdot 23$
$\binom{28}{14} = 2^3\cdot 3^3\cdot 5^2\cdot 17\cdot 19\cdot 23$
$\binom{30}{15} = 2^4\cdot 3^2\cdot 5\cdot 17\cdot 19\cdot 23\cdot 29$
mostrano che fattori primi "molto piccoli" $p < \sqrt{2n}$ possono apparire come potenze più grandi in $\binom{2n}{n}$, "piccoli" numeri primi con $\sqrt{2n} < p \leq \frac{2}{3}n$ appaiono al più una sola volta, mentre fattori nell'intervallo con $\frac{2}{3}n < p \leq n$ non appaiono per nulla

e pertanto per $n \geq 50$ (e quindi $18 < 2\sqrt{2n}$) otteniamo da (3) e (4)

$$2^{2n} \leq (2n)^{3\left(1+\sqrt{2n}\right)} < 2^{\sqrt[6]{2n}\left(18+18\sqrt{2n}\right)} < 2^{20\sqrt[6]{2n}\sqrt{2n}} = 2^{20(2n)^{2/3}}.$$

Ciò implica $(2n)^{1/3} < 20$, e dunque $n < 4000$. □

Si può dedurre anche di più da questo genere di stime: dalla (2) possiamo derivare con gli stessi metodi che

$$\prod_{n<p\leq 2n} p \geq 2^{\frac{1}{30}n} \quad \text{per} \quad n \geq 4000,$$

e dunque che ci sono almeno

$$\log_{2n}\left(2^{\frac{1}{30}n}\right) \;=\; \frac{1}{30}\,\frac{n}{\log_2 n + 1}$$

numeri primi compresi fra n e $2n$.

Questa stima non è poi tanto grossolana: il "vero" numero di numeri primi in questo intervallo è all'incirca $n/\log n$. Ciò deriva dal "teorema dei numeri primi", il quale afferma che il limite

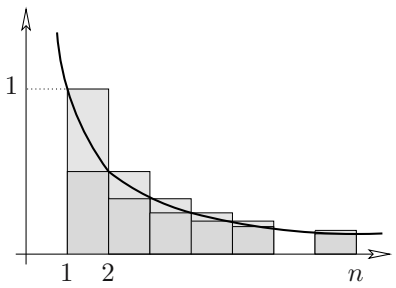
$$\lim_{n\to\infty} \frac{\#\{p \leq n : p \text{ è primo}\}}{n/\log n}$$

esiste, ed è uguale a 1. Questo è un risultato famoso che fu dimostrato per la prima volta da Hadamard e de la Vallée-Poussin nel 1896; Selberg e Erdős trovarono una dimostrazione elementare nel 1948, che pur non facendo uso di tecniche di analisi complessa, è lunga e laboriosa.

Sembra tuttavia che sul teorema dei numeri primi non sia ancora stata detta l'ultima parola: ad esempio la dimostrazione dell'ipotesi di Riemann (si veda a pagina 47), uno dei principali problemi aperti, tuttora irrisolti, della matematica, fornirebbe anche un miglioramento sostanziale della stima del teorema dei numeri primi. Ma ne deriverebbe un decisivo miglioramento anche per il postulato di Bertrand. In effetti, quello che segue è un famoso problema irrisolto:

> *È sempre possibile trovare un numero primo compreso fra n^2 e $(n+1)^2$?*

Per ulteriori informazioni si vedano [3, p. 19] e [4, pp. 248, 257].



Appendice: alcune stime

Stime attraverso integrali

Esiste un metodo molto semplice ma efficace per stimare somme utilizzando integrali (come già visto a pagina 4). Per stimare *i numeri armonici*

$$H_n \;=\; \sum_{k=1}^{n} \frac{1}{k},$$

tracciamo la figura a margine della pagina precedente. Confrontando l'area al di sotto del grafico di $f(t) = \frac{1}{t}$ $(1 \leq t \leq n)$ con l'area dei rettangoli ombreggiati scuri deriviamo che

$$H_n - 1 \;=\; \sum_{k=2}^{n} \frac{1}{k} \;<\; \int_1^n \frac{1}{t}\, dt \;=\; \log n,$$

mentre, confrontandola con l'area dei rettangoli grandi (che includono le parti ombreggiate chiare), otteniamo che

$$H_n - \frac{1}{n} \;=\; \sum_{k=1}^{n-1} \frac{1}{k} \;>\; \int_1^n \frac{1}{t}\, dt \;=\; \log n.$$

Considerando entrambe le disuguaglianze, otteniamo

$$\log n + \frac{1}{n} \;<\; H_n \;<\; \log n \;+\; 1.$$

In particolare, $\lim_{n\to\infty} H_n = \infty$, e l'ordine di infinito di H_n è dato da $\lim_{n\to\infty} \frac{H_n}{\log n} = 1$. Tuttavia si conoscono stime decisamente migliori (si veda [2]), come ad esempio

$$H_n \;=\; \log n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + O\left(\frac{1}{n^6}\right),$$

Qui $O\left(\frac{1}{n^6}\right)$ indica una funzione $f(n)$ tale che la disuguaglianza $f(n) \leq c\frac{1}{n^6}$ sia vera per una opportuna costante c

essendo $\gamma \approx 0.5772$ la "costante di Eulero".

Stima dei fattoriali — formula di Stirling

Lo stesso metodo applicato a

$$\log(n!) \;=\; \log 2 + \log 3 + \ldots + \log n \;=\; \sum_{k=2}^{n} \log k$$

fornisce

$$\log((n-1)!) \;<\; \int_1^n \log t\, dt \;<\; \log(n!),$$

dove l'integrale è facilmente calcolabile, essendo

$$\int_1^n \log t\, dt \;=\; \Big[t \log t - t\Big]_1^n \;=\; n \log n - n + 1.$$

Pertanto otteniamo una minorazione per $n!$

$$n! \;>\; e^{n \log n - n + 1} \;=\; e\Big(\frac{n}{e}\Big)^n$$

e allo stesso tempo una maggiorazione

$$n! \;=\; n\,(n-1)! \;<\; n e^{n \log n - n + 1} \;=\; en\Big(\frac{n}{e}\Big)^n.$$

Un'analisi più attenta è necessaria in questo caso per descrivere il comportamento asintotico di $n!$, come si ottiene dalla *formula di Stirling*

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Esistono versioni ancora più precise, come ad esempio

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1+\frac{1}{12n}+\frac{1}{288n^2}-\frac{139}{5140n^3}+O\left(\frac{1}{n^4}\right)\right).$$

Qui $f(n) \sim g(n)$ significa che $\lim_{n\to\infty}\frac{f(n)}{g(n)} = 1$

Stima di coefficienti binomiali

Proprio dalla definizione dei coefficienti binomiali $\binom{n}{k}$ come numero di k-sottoinsiemi di un n-insieme, sappiamo che la successione $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$ di coefficienti binomiali

- ha come somma $\sum_{k=0}^{n}\binom{n}{k} = 2^n$
- è simmetrica: $\binom{n}{k} = \binom{n}{n-k}$.

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
 & 7 & & 21 & & 35 & & 35 & & 21 & & 7 & \\
\end{array}$$

Il triangolo di Pascal

Dall'equazione funzionale $\binom{n}{k} = \frac{n-k+1}{k}\binom{n}{k-1}$ si ottiene facilmente che per ogni n i coefficienti binomiali $\binom{n}{k}$ formano una successione simmetrica e *unimodale*: essa cresce verso il centro, di modo che i coefficienti binomiali centrali sono i più grandi della successione

$$1 = \binom{n}{0} < \binom{n}{1} < \ldots < \binom{n}{\lfloor n/2\rfloor} = \binom{n}{\lceil n/2\rceil} > \ldots > \binom{n}{n-1} > \binom{n}{n} = 1.$$

Qui $\lfloor x\rfloor$ risp. $\lceil x\rceil$ indica il numero x arrotondato per difetto risp. per eccesso al più vicino intero.

Dalle formule asintotiche per i fattoriali precedentemente citate si possono ottenere stime molto precise per le grandezze dei coefficienti binomiali. Tuttavia, in questo libro faremo uso solo di stime semplici e molto deboli, quali ad esempio: $\binom{n}{k} \le 2^n$ per ogni k, mentre per $n \ge 2$ abbiamo

$$\binom{n}{\lfloor n/2\rfloor} \ge \frac{2^n}{n},$$

essendo l'uguaglianza valida solo per $n = 2$. In particolare, per $n \ge 1$,

$$\binom{2n}{n} \ge \frac{4^n}{2n}.$$

Questo vale poiché $\binom{n}{\lfloor n/2\rfloor}$, un coefficiente binomiale mediano, è il più grande elemento della successione $\binom{n}{0}+\binom{n}{n}, \binom{n}{1}, \binom{n}{2}, \ldots, \binom{n}{n-1}$, la cui somma vale 2^n, e la cui media è pertanto $\frac{2^n}{n}$.

D'altro canto, per i coefficienti binomiali abbiamo la maggiorazione:

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \le \frac{n^k}{k!} \le \frac{n^k}{2^{k-1}},$$

che fornisce una stima ragionevolmente buona per i coefficienti binomiali "piccoli" nella coda della successione, quando n è grande (rispetto a k).

Bibliografia

[1] P. ERDŐS: *Beweis eines Satzes von Tschebyschef,* Acta Sci. Math. (Szeged) **5** (1930-32), 194-198.

[2] R. L. GRAHAM, D. E. KNUTH & O. PATASHNIK: *Concrete Mathematics. A Foundation for Computer Science,* Addison-Wesley, Reading MA 1989.

[3] G. H. HARDY & E. M. WRIGHT: *An Introduction to the Theory of Numbers,* fifth edition, Oxford University Press 1979.

[4] P. RIBENBOIM: *The New Book of Prime Number Records,* Springer-Verlag, New York 1989.

I coefficienti binomiali non sono (quasi) mai potenze

Capitolo 3

Esiste un epilogo al postulato di Bertrand che fornisce uno splendido risultato sui coefficienti binomiali. Nel 1892 Sylvester rafforzò il postulato di Bertrand nel seguente modo:

> *Se $n \geq 2k$, allora almeno uno dei numeri $n, n-1, \ldots, n-k+1$ ha un divisore primo p più grande di k.*

Si noti che per $n = 2k$ si ottiene precisamente il postulato di Bertrand. Nel 1934, Erdős fornì una *Book Proof* breve ed elementare del risultato di Sylvester, seguendo la traccia della sua dimostrazione del postulato di Bertrand. Esiste un enunciato equivalente per il teorema di Sylvester:

> *Il coefficiente binomiale*
>
> $$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} \qquad (n \geq 2k)$$
>
> *ha sempre un fattore primo $p > k$.*

Tenendo a mente questa osservazione, ci interessiamo ad un altro dei gioielli di Erdős. Quando $\binom{n}{k}$ è uguale ad una potenza m^ℓ? È semplice notare che ci sono infinite soluzioni per $k = \ell = 2$, ovvero soluzioni dell'equazione $\binom{n}{2} = m^2$. In effetti, se $\binom{n}{2}$ è un quadrato, allora lo è anche $\binom{(2n-1)^2}{2}$. Per verificarlo, si ponga $n(n-1) = 2m^2$. Ne segue che

$$(2n-1)^2((2n-1)^2-1) = (2n-1)^2 4n(n-1) = 2(2m(2n-1))^2,$$

e pertanto

$$\binom{(2n-1)^2}{2} = (2m(2n-1))^2.$$

A cominciare da $\binom{9}{2} = 6^2$ otteniamo dunque infinite soluzioni — la successiva è $\binom{289}{2} = 204^2$. Ciò tuttavia non fornisce tutte le soluzioni. Ad esempio, $\binom{50}{2} = 35^2$ inizia un'altra serie, e così pure $\binom{1682}{2} = 1189^2$. Per $k = 3$ è noto che $\binom{n}{3} = m^2$ ha come unica soluzione $n = 50$, $m = 140$. Ma ora siamo giunti alla fine. Per $k \geq 4$ ed un qualsiasi $\ell \geq 2$ non esiste alcuna soluzione, come dimostrò Erdős con un'ingegnosa argomentazione.

$\binom{50}{3} = 140^2$
è l'unica soluzione per $k = 3$, $\ell = 2$

Teorema. *L'equazione $\binom{n}{k} = m^\ell$ non ha soluzioni intere per $\ell \geq 2$ e $4 \leq k \leq n-4$.*

■ **Dimostrazione.** Si noti dapprima che possiamo supporre $n \geq 2k$ dal momento che $\binom{n}{k} = \binom{n}{n-k}$. Supponiamo che il teorema sia falso, e che $\binom{n}{k} = m^\ell$. La dimostrazione, per contraddizione, si articola nelle quattro fasi seguenti.

(1) Secondo il teorema di Sylvester, esiste un fattore primo p di $\binom{n}{k}$ maggiore di k, pertanto p^ℓ divide $n(n-1)\cdots(n-k+1)$. Chiaramente, soltanto uno dei fattori $n-i$ può essere un multiplo di p (essendo $p > k$); possiamo dunque concludere che $p^\ell \mid n-i$ e pertanto

$$n \;\geq\; p^\ell \;>\; k^\ell \;\geq\; k^2.$$

(2) Si consideri un qualunque fattore $n-j$ del numeratore e lo si scriva nella forma $n-j = a_j m_j^\ell$, dove a_j non è divisibile da alcuna potenza ℓ-esima non banale. Grazie a **(1)**, notiamo che a_j ha unicamente divisori primi inferiori o uguali a k. Vogliamo quindi mostrare che $a_i \neq a_j$ se $i \neq j$. Supponiamo al contrario che $a_i = a_j$ per qualche $i < j$. Allora $m_i \geq m_j + 1$ e

$$\begin{aligned} k \;&>\; (n-i)-(n-j) \;=\; a_j(m_i^\ell - m_j^\ell) \;\geq\; a_j((m_j+1)^\ell - m_j^\ell) \\ &>\; a_j \ell m_j^{\ell-1} \;\geq\; \ell(a_j m_j^\ell)^{1/2} \;\geq\; \ell(n-k+1)^{1/2} \\ &\geq\; \ell(\tfrac{n}{2}+1)^{1/2} \;>\; n^{1/2}, \end{aligned}$$

il che contraddice la disuguaglianza $n > k^2$ riportata sopra.

(3) Dimostriamo ora che gli a_i sono gli interi $1, 2, \ldots, k$ in un ordine dato. Secondo Erdős, questo è il punto cruciale della dimostrazione. Poiché sappiamo già che tali numeri sono tutti distinti, è sufficiente dimostrare che

$$a_0 a_1 \cdots a_{k-1} \quad \text{divide} \quad k!.$$

Sostituendo $n-j = a_j m_j^\ell$ nell'equazione $\binom{n}{k} = m^\ell$, otteniamo

$$a_0 a_1 \cdots a_{k-1}(m_0 m_1 \cdots m_{k-1})^\ell \;=\; k! m^\ell.$$

Eliminando i fattori comuni di $m_0 \cdots m_{k-1}$ e m si trova

$$a_0 a_1 \cdots a_{k-1} u^\ell \;=\; k! v^\ell$$

con $\text{mcd}(u, v) = 1$. Resta da dimostrare che $v = 1$. Se così non fosse, v conterrebbe un divisore primo p. Poiché $\text{mcd}(u, v) = 1$, p deve essere un divisore primo di $a_0 a_1 \cdots a_{k-1}$ e pertanto è inferiore o uguale a k. Dal teorema di Legendre (si veda a pagina 2) sappiamo che $k!$ contiene p alla potenza $\sum_{i \geq 1} \lfloor \frac{k}{p^i} \rfloor$. Stimiamo ora l'esponente di p in $n(n-1)\cdots(n-k+1)$. Siano i un intero positivo e $b_1 < b_2 < \ldots < b_s$ i multipli di p^i tra $n, n-1, \ldots, n-k+1$. Allora $b_s = b_1 + (s-1)p^i$ e dunque

$$(s-1)p^i \;=\; b_s - b_1 \;\leq\; n - (n-k+1) \;=\; k-1,$$

il che implica

$$s \;\leq\; \left\lfloor \frac{k-1}{p^i} \right\rfloor + 1 \;\leq\; \left\lfloor \frac{k}{p^i} \right\rfloor + 1.$$

Pertanto, per ogni i il numero di multipli di p^i tra $n, \ldots, n-k+1$, e quindi tra gli a_j, è limitato da $\lfloor \frac{k}{p^i} \rfloor + 1$. Ragionando come fatto con il teorema di Legendre nel Capitolo 2, ciò implica che l'esponente di p in $a_0 a_1 \cdots a_{k-1}$ vale al più

$$\sum_{i=1}^{\ell-1} \left(\left\lfloor \frac{k}{p^i} \right\rfloor + 1 \right).$$

L'unica differenza è che questa volta la somma si arresta ad $i = \ell - 1$, in quanto gli a_j non contengono alcuna potenza ℓ-esima. Unendo questi due elementi, troviamo che l'esponente di p in v^ℓ vale al massimo

$$\sum_{i=1}^{\ell-1} \left(\left\lfloor \frac{k}{p^i} \right\rfloor + 1 \right) - \sum_{i \geq 1} \left\lfloor \frac{k}{p^i} \right\rfloor \;\leq\; \ell - 1,$$

ottenendo così la contraddizione desiderata, poiché v^ℓ è una potenza ℓ-esima.

Ciò è sufficiente a sistemare il caso $\ell = 2$. In effetti, dal momento che $k \geq 4$, uno degli a_i deve essere uguale a 4, ma gli a_i non contengono quadrati. Passiamo a trattare il caso $\ell \geq 3$.

Notiamo che la nostra analisi finora è in accordo con $\binom{50}{3} = 140^2$, essendo

$$\begin{aligned} 50 &= 2 \cdot 5^2 \\ 49 &= 1 \cdot 7^2 \\ 48 &= 3 \cdot 4^2 \end{aligned}$$

e $5 \cdot 7 \cdot 4 = 140$

(4) Poiché $k \geq 4$, dobbiamo avere $a_{i_1} = 1$, $a_{i_2} = 2$, $a_{i_3} = 4$ per qualche i_1, i_2, i_3, ovvero,

$$n - i_1 = m_1^\ell, \;\; n - i_2 = 2m_2^\ell, \;\; n - i_3 = 4m_3^\ell.$$

Affermiamo che $(n-i_2)^2 \neq (n-i_1)(n-i_3)$. Altrimenti, posto $b = n - i_2$ e $n - i_1 = b - x$, $n - i_3 = b + y$, dove $0 < |x|, |y| < k$, ne deriva che

$$b^2 = (b-x)(b+y) \quad \text{oppure} \quad (y-x)b = xy,$$

dove $x = y$ è chiaramente impossibile. Ora da **(1)** segue

$$|xy| \;=\; b|y-x| \;\geq\; b \;>\; n-k \;>\; (k-1)^2 \;\geq\; |xy|,$$

il che è assurdo.

Pertanto abbiamo che $m_2^2 \neq m_1 m_3$, dove supponiamo che $m_2^2 > m_1 m_3$ (l'altro caso essendo analogo), e procediamo alle ultime catene di diseguaglianze. Otteniamo che

$$\begin{aligned} 2(k-1)n \;&>\; n^2 - (n-k+1)^2 \;>\; (n-i_2)^2 - (n-i_1)(n-i_3) \\ &=\; 4[m_2^{2\ell} - (m_1 m_3)^\ell] \;\geq\; 4[(m_1 m_3 + 1)^\ell - (m_1 m_3)^\ell] \\ &\geq\; 4\ell m_1^{\ell-1} m_3^{\ell-1}. \end{aligned}$$

Poiché $\ell \geq 3$ e $n > k^\ell \geq k^3 > 6k$, ne deriva che

$$\begin{aligned} 2(k-1)n m_1 m_3 \;&>\; 4\ell m_1^\ell m_3^\ell \;=\; \ell(n-i_1)(n-i_3) \\ &>\; \ell(n-k+1)^2 \;>\; 3(n - \frac{n}{6})^2 \;>\; 2n^2. \end{aligned}$$

Ora, poiché $m_i \leq n^{1/\ell} \leq n^{1/3}$ concludiamo che

$$kn^{2/3} \ \geq \ km_1m_3 \ > \ (k-1)m_1m_3 \ > \ n,$$

ovvero $k^3 > n$. Questa contraddizione completa la dimostrazione. □

Bibliografia

[1] P. ERDŐS: *A theorem of Sylvester and Schur,* J. London Math. Soc. **9** (1934), 282-288.

[2] P. ERDŐS: *On a diophantine equation,* J. London Math. Soc. **26** (1951), 176-178.

[3] J. J. SYLVESTER: *On arithmetical series,* Messenger of Math. **21** (1892), 1-19, 87-120; Collected Mathematical Papers Vol. 4, 1912, 687-731.

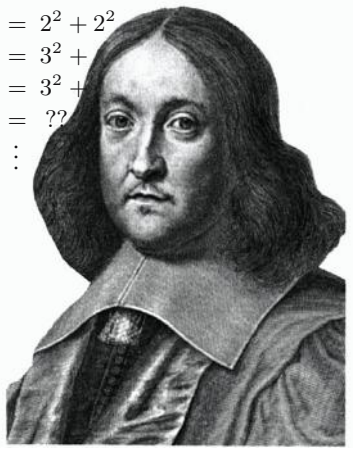
Rappresentazione di numeri come somme di due quadrati

Capitolo 4

Quali numeri possono essere scritti come somme di due quadrati?

Questa domanda è vecchia come la teoria dei numeri e la sua soluzione è un classico in questo campo. La parte "difficile" della soluzione è controllare che ogni numero primo della forma $4m + 1$ è la somma di due quadrati. G. H. Hardy scrive che questo *teorema dei due quadrati* di Fermat "è considerato, molto giustamente, come uno dei più raffinati dell'aritmetica". Ciononostante, una delle nostre *Book Proofs* riportata di seguito è piuttosto recente.

$$\begin{aligned} 1 &= 1^2 + 0^2 \\ 2 &= 1^2 + 1^2 \\ 3 &= \ ?? \\ 4 &= 2^2 + 0^2 \\ 5 &= 2^2 + 1^2 \\ 6 &= \ ?? \\ 7 &= \ ?? \\ 8 &= 2^2 + 2^2 \\ 9 &= 3^2 + \\ 10 &= 3^2 + \\ 11 &= \ ?? \\ &\vdots \end{aligned}$$



Pierre de Fermat

Cominciamo con un po' di "riscaldamento". Dapprima, dobbiamo distinguere tra il numero primo $p = 2$, i numeri primi della forma $p = 4m+1$ ed i numeri primi della forma $p = 4m + 3$. Ogni numero primo appartiene ad una ed una sola di queste tre classi. A questo punto possiamo notare (usando un metodo "alla Euclide") che esistono infiniti numeri primi della forma $4m + 3$. Infatti, se ne esistesse soltanto un numero finito, allora potremmo prendere p_k come il più grande di questa forma. Ponendo

$$N_k \ := \ 2^2 \cdot 3 \cdot 5 \cdots p_k - 1$$

(dove $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, ... rappresenta la successione di tutti i numeri primi), troviamo che N_k è congruente a $3 \pmod 4$, quindi deve avere un fattore primo della forma $4m + 3$; poiché tale fattore primo è maggiore di p_k si ha una contraddizione. Al termine di questo capitolo dedurremo anche che esiste un numero infinito di numeri primi dell'altro tipo, $p = 4m + 1$.

Il nostro primo lemma è un caso particolare della celebre "legge della reciprocità": essa caratterizza i numeri primi per i quali -1 è un quadrato nel campo $\mathbb{Z}_p$ (si veda il riquadro nella prossima pagina).

Lemma 1. *Per i numeri primi $p = 4m + 1$ l'equazione $s^2 \equiv -1 \pmod p$ ha due soluzioni $s \in \{1, 2, \ldots, p-1\}$, per $p = 2$ ne ha una, mentre per i numeri primi della forma $p = 4m + 3$ non ne ha alcuna.*

■ **Dimostrazione.** Per $p = 2$ si prenda $s = 1$. Per i p dispari, costruiamo la relazione d'equivalenza su $\{1, 2, \ldots, p - 1\}$ generata identificando ogni elemento con il proprio additivo inverso e con il proprio moltiplicativo inverso in $\mathbb{Z}_p$. Pertanto le classi di equivalenza "generali" conterranno quattro elementi

$$\{x, -x, \overline{x}, -\overline{x}\}$$

poiché tale insieme di 4 elementi contiene entrambi gli inversi di ciascuno dei suoi elementi. Tuttavia, ci sono classi d'equivalenza più piccole se qualcuno dei quattro numeri non è distinto:

- $x \equiv -x$ è impossibile per p dispari.
- $x \equiv \overline{x}$ è equivalente a $x^2 \equiv 1$. Vi sono due soluzioni, $x = 1$ e $x = p-1$, che conducono alla classe d'equivalenza $\{1, p-1\}$ di dimensione 2.
- $x \equiv -\overline{x}$ è equivalente a $x^2 \equiv -1$. Questa equazione può non avere soluzione oppure avere due soluzioni distinte $x_0, p - x_0$: in tal caso la classe d'equivalenza è $\{x_0, p - x_0\}$.

Per $p = 11$ la partizione è $\{1, 10\}, \{2, 9, 6, 5\}, \{3, 8, 4, 7\}$; per $p = 13$ è $\{1, 12\}, \{2, 11, 7, 6\}, \{3, 10, 9, 4\}, \{5, 8\}$: la coppia $\{5, 8\}$ fornisce le due soluzioni di $s^2 \equiv -1 \bmod 13$

L'insieme $\{1, 2, \ldots, p-1\}$ ha $p-1$ elementi; abbiamo effettuato delle partizioni in quadruple (classi di equivalenza di dimensione 4) ed in una o due coppie (classi di equivalenza di dimensione 2). Per $p - 1 = 4m + 2$ troviamo che vi è una sola coppia $\{1, p-1\}$, le restanti partizioni sono quadruple; perciò $s^2 \equiv -1 \pmod p$ non ha soluzione. Per $p - 1 = 4m$ ci deve essere una seconda coppia, e questa contiene le due soluzioni di $s^2 \equiv -1$ che cercavamo. □

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\cdot$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Addizione e moltiplicazione in $\mathbb{Z}_5$

Campi primi

Se p è un numero primo, allora l'insieme $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$ con addizione e moltiplicazione definite "modulo p" forma un campo finito. Avremo bisogno delle seguenti proprietà:

- Per $x \in \mathbb{Z}_p$, $x \neq 0$, l'inverso additivo (per il quale si usa la notazione $-x$) è dato da $p - x \in \{1, 2, \ldots, p-1\}$. Se $p > 2$, allora x e $-x$ sono elementi distinti di $\mathbb{Z}_p$.
- Ogni $x \in \mathbb{Z}_p \backslash \{0\}$ ha un unico fattore inverso $\overline{x} \in \mathbb{Z}_p \backslash \{0\}$, tale che $x\overline{x} \equiv 1 \pmod p$.
  La definizione dei numeri primi implica che la corrispondenza $\mathbb{Z}_p \to \mathbb{Z}_p$, $z \mapsto xz$ è iniettiva per $x \neq 0$. Pertanto sull'insieme finito $\mathbb{Z}_p \backslash \{0\}$ essa deve essere anche suriettiva; dunque per ogni x esiste un unico $\overline{x} \neq 0$ tale che $x\overline{x} \equiv 1 \pmod p$.
- I quadrati $0^2, 1^2, 2^2, \ldots, h^2$ definiscono elementi diversi di $\mathbb{Z}_p$, con $h = \lfloor \frac{p}{2} \rfloor$.
  In effetti $x^2 \equiv y^2$, ovvero $(x + y)(x - y) \equiv 0$, implica che $x \equiv y$ oppure che $x \equiv -y$. Gli $1 + \lfloor \frac{p}{2} \rfloor$ elementi $0^2, 1^2, \ldots, h^2$ sono detti i *quadrati* in $\mathbb{Z}_p$.

A questo punto, osserviamo "al volo" che per *tutti* i numeri primi vi sono soluzioni per $x^2 + y^2 \equiv -1 \pmod p$. Vi sono infatti $\lfloor \frac{p}{2} \rfloor + 1$ quadrati distinti x^2 in $\mathbb{Z}_p$, ed esistono $\lfloor \frac{p}{2} \rfloor + 1$ numeri distinti della forma $-(1 + y^2)$.

Questi due insiemi di numeri sono troppo grandi per essere disgiunti, dal momento che $\mathbb{Z}_p$ ha solo p elementi; pertanto devono esistere x ed y tali che $x^2 \equiv -(1+y^2) \pmod p$.

Lemma 2. *Nessun numero $n = 4m+3$ è somma di due quadrati.*

■ **Dimostrazione.** Il quadrato di ogni numero pari è $(2k)^2 = 4k^2 \equiv 0 \pmod 4$, mentre i quadrati di numeri dispari danno $(2k+1)^2 = 4(k^2+k)+1 \equiv 1 \pmod 4$. Pertanto ogni somma di due quadrati è congruente a $0, 1$ o $2 \pmod 4$. □

Ci sembra pertanto evidente che i numeri primi $p = 4m+3$ non vanno bene. Pertanto procediamo con le "buone" proprietà di cui godono i numeri primi della forma $p = 4m+1$. Il passaggio seguente è quello fondamentale per giungere al teorema principale.

Proposizione. *Ogni numero primo della forma $p = 4m+1$ è somma di due quadrati, ovvero può essere scritto come $p = x^2+y^2$, essendo $x, y \in \mathbb{N}$ due opportuni numeri naturali.*

Presenteremo due dimostrazioni di questo risultato — entrambe eleganti e sorprendenti. La prima è caratterizzata da un'applicazione notevole del principio del casellario (che abbiamo già usato "al volo" prima del Lemma 2; si veda il capitolo 22 per ulteriori dettagli), così come di una mossa intelligente verso e da argomenti "modulo p". L'idea risale al teorico dei numeri norvegese Axel Thue.

■ **Dimostrazione.** Si considerino le coppie di interi (x', y') con $0 \le x', y' \le \sqrt{p}$, ovvero, $x', y' \in \{0, 1, \ldots, \lfloor\sqrt{p}\rfloor\}$. Ve ne sono esattamente $(\lfloor\sqrt{p}\rfloor + 1)^2$. Usando la stima $\lfloor x\rfloor + 1 > x$ per $x = \sqrt{p}$, vediamo che abbiamo più di p coppie di interi di questo tipo. Pertanto per ogni $s \in \mathbb{Z}$, è impossibile che tutti i valori $x' - sy'$ prodotti dalle coppie (x', y') siano distinti modulo p. Ciò significa che per ogni s ci sono due coppie distinte

$$(x', y'),\ (x'', y'') \ \in\ \{0, 1, \ldots, \lfloor\sqrt{p}\rfloor\}^2$$

con

$$x' - sy' \ \equiv\ x'' - sy'' \pmod p.$$

Prendiamo ora le differenze: abbiamo $x' - x'' \equiv s(y' - y'') \pmod p$. Pertanto se definiamo

$$x := |x' - x''|, \quad y := |y' - y''|,$$

otteniamo

$$(x, y) \in \{0, 1, \ldots, \lfloor\sqrt{p}\rfloor\}^2 \quad \text{con} \quad x \equiv \pm sy \pmod p.$$

Sappiamo inoltre che non è possibile che entrambi x e y valgano zero, essendo le coppie (x', y') e (x'', y'') distinte.

Sia ora s una soluzione di $s^2 \equiv -1 \pmod p$; essa esiste in virtù del Lemma 1. Allora $x^2 \equiv s^2y^2 \equiv -y^2 \pmod p$; abbiamo così ottenuto

$$(x, y) \in \mathbb{Z}^2 \quad \text{con} \quad 0 < x^2 + y^2 < 2p \quad \text{e} \quad x^2 + y^2 \equiv 0 \pmod p.$$

Per $p = 13$, $\lfloor\sqrt{p}\rfloor = 3$ consideriamo $x', y' \in \{0, 1, 2, 3\}$. Per $s = 5$, la somma $x' - sy' \pmod{13}$ assume i seguenti valori:

x' \ y'	0	1	2	3
0	0	8	3	11
1	1	9	4	12
2	2	10	5	0
3	3	11	6	1

Ma p è l'unico numero tra 0 e $2p$ ad essere divisibile per p. Pertanto $x^2 + y^2 = p$: ed abbiamo finito! □

La nostra seconda dimostrazione della proposizione — chiaramente anch'essa una *Book Proof* — fu scoperta da Roger Heath-Brown nel 1971 ed apparve nel 1984. (Una versione condensata in una singola frase ne fu data da Don Zagier.) Essa è talmente elementare che non abbiamo neppure bisogno del Lemma 1.

L'argomentazione di Heath-Brown si compone di tre involuzioni lineari: una alquanto ovvia, una nascosta ed una banale che dà il *coup de grâce*. La seconda ed inattesa involuzione corrisponde ad una struttura nascosta nell'insieme delle soluzioni intere dell'equazione $4xy + z^2 = p$.

■ **Dimostrazione.** Studiamo l'insieme

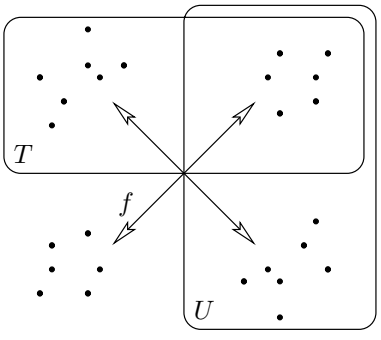
$$S \;:=\; \{(x, y, z) \in \mathbb{Z}^3 : 4xy + z^2 = p, \quad x > 0, \quad y > 0\}.$$

Tale insieme è finito. In effetti, $x \geq 1$ e $y \geq 1$ implica che $y \leq \frac{p}{4}$ e $x \leq \frac{p}{4}$. Vi è pertanto solo un numero finito di valori possibili per x e y, e dati x e y, vi sono al più due valori per z.

1. La prima involuzione lineare è data da

$$f : S \longrightarrow S, \quad (x, y, z) \longmapsto (y, x, -z),$$

ossia, "scambiare x e y, e negare z." Ciò ovviamente trasforma S in se stesso ed è pertanto un'*involuzione*: applicata due volte, dà l'identità. Inoltre, f non ha punti fissi, poiché $z = 0$ implicherebbe $p = 4xy$, il che è impossibile. Inoltre, f trasforma le soluzioni in



$$T \;:=\; \{(x, y, z) \in S : z > 0\}$$

nelle soluzioni in $S \backslash T$ che soddisfano $z < 0$. Si aggiunga che f inverte i segni di $x - y$ e di z, e dunque trasforma le soluzioni in

$$U \;:=\; \{(x, y, z) \in S : (x - y) + z > 0\}$$

nelle soluzioni in $S \backslash U$. Per questo dobbiamo constatare che non vi sono soluzioni tali che $(x - y) + z = 0$ e che non ve ne è nessuna perché ciò darebbe $p = 4xy + z^2 = 4xy + (x - y)^2 = (x + y)^2$.

Che cosa deduciamo dallo studio di f? L'osservazione principale è che poiché f trasforma gli insiemi T e U nei loro complementi, essa scambia anche gli elementi di $T \backslash U$ con quelli di $U \backslash T$. Ciò significa che vi è lo stesso numero di soluzioni in U non appartenenti a T che di soluzioni in T ma non appartenenti a U — *pertanto T e U hanno la stessa cardinalità*.

2. La seconda involuzione che studiamo è un'involuzione sull'insieme U:

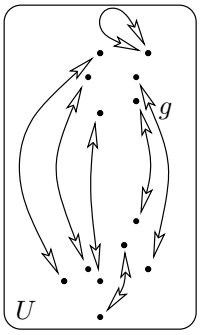
$$g : U \longrightarrow U, \quad (x, y, z) \longmapsto (x - y + z, y, 2y - z).$$

Innanzitutto controlliamo che questa sia effettivamente una corrispondenza ben definita: se $(x, y, z) \in U$, allora $x - y + z > 0, y > 0$ e $4(x - y +$

$z)y + (2y - z)^2 = 4xy + z^2$, dunque $g(x, y, z) \in S$. Usando $(x - y + z) - y + (2y - z) = x > 0$ troviamo effettivamente che $g(x, y, z) \in U$.
Inoltre, g è un'involuzione, in quanto $g(x, y, z) = (x - y + z, y, 2y - z)$ è trasformata da g su $((x-y+z)-y+(2y-z), y, 2y-(2y-z)) = (x, y, z)$.
Infine, g ha esattamente un punto fisso:

$$(x, y, z) \;=\; g(x, y, z) \;=\; (x - y + z, y, 2y - z)$$

vale esattamente se $y = z$; ma allora $p = 4xy + y^2 = (4x + y)y$, il che vale solo per $y = 1 = z$ e $x = \frac{p-1}{4}$.
Ma se g è un'involuzione su U che ha esattamente un punto fisso, allora *la cardinalità di U è dispari*.

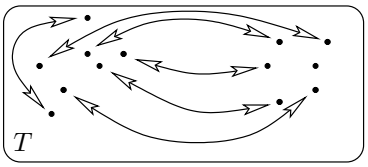


3. La terza, banale, involuzione che studiamo è l'involuzione su T che scambia x e y:

$$h : T \longrightarrow T, \quad (x, y, z) \longmapsto (y, x, z).$$

Questa corrispondenza è chiaramente ben definita ed è un'involuzione.
Combiniamo ora quanto abbiamo appreso dalle due altre involuzioni: la cardinalità di T è uguale alla cardinalità di U, che è dispari. Ma se h è un'involuzione su un insieme finito di cardinalità dispari, essa *ha un punto fisso*: vi è un punto $(x, y, z) \in T$ tale che $x = y$, ovvero, una soluzione di

$$p = 4x^2 + z^2 = (2x)^2 + z^2. \qquad \square$$



Su un insieme finito di cardinalità dispari, ogni involuzione ha almeno un punto fisso

Si noti che questa dimostrazione dà un risultato più preciso: il numero di rappresentazioni di p nella forma $p = x^2 + (2y)^2$ è *dispari* per tutti i numeri primi della forma $p = 4m + 1$. La rappresentazione è in effetti unica: si veda [3]. Si noti anche che entrambe le dimostrazioni sono poco efficaci: si provi a trovare x e y per un numero primo di dieci cifre! Metodi efficaci per trovare tali rappresentazioni come somme di due quadrati sono discussi in [1] e [7].
Il seguente teorema dà una risposta completa alla domanda con cui è iniziato questo capitolo.

Teorema. *Un numero naturale n può essere rappresentato come somma di due quadrati se e solo se ogni fattore primo della forma $p = 4m + 3$ appare con un esponente pari nella decomposizione in primi di n.*

■ **Dimostrazione.** Diciamo che un numero n è *rappresentabile* se esso è somma di due quadrati, ovvero, se $n = x^2 + y^2$ per qualche $x, y \in \mathbb{N}_0$. Il teorema è una conseguenza dei seguenti cinque fatti.

(1) $1 = 1^2 + 0^2$ e $2 = 1^2 + 1^2$ sono rappresentabili. Ogni numero primo della forma $p = 4m + 1$ è rappresentabile.

(2) Il prodotto di due numeri rappresentabili qualsiasi $n_1 = x_1^2 + y_1^2$ e $n_2 = x_2^2 + y_2^2$ è rappresentabile: $n_1 n_2 = (x_1 x_2 + y_1 y_2)^2 + (x_1 y_2 - x_2 y_1)^2$.

(3) Se n è rappresentabile, $n = x^2+y^2$, allora anche nz^2 è rappresentabile, poiché $nz^2 = (xz)^2 + (yz)^2$.

I fatti (1), (2) e (3) insieme costituiscono l'ipotesi del teorema.

(4) Se $p = 4m + 3$ è un numero primo che divide un numero rappresentabile $n = x^2 + y^2$, allora p divide sia x sia y; pertanto p^2 divide n. In effetti, se avessimo $x \not\equiv 0 \pmod p$, potremmo allora trovare $\overline{x}$ tale che $x\overline{x} \equiv 1 \pmod p$, moltiplicare l'equazione $x^2 + y^2 \equiv 0$ per $\overline{x}^2$, e quindi ottenere $1 + y^2\overline{x}^2 = 1 + (\overline{x}y)^2 \equiv 0 \pmod p$, il che è impossibile per $p = 4m + 3$ per via del Lemma 1.

(5) Se n è rappresentabile e $p = 4m+3$ divide n, allora p^2 divide n, e n/p^2 è rappresentabile. Ciò segue da (4) e completa la dimostrazione. □

Come corollario, possiamo concludere che esiste un numero infinito di numeri primi della forma $p = 4m + 1$. A tal fine, consideriamo

$$M_k = (3 \cdot 5 \cdot 7 \cdots p_k)^2 + 2^2,$$

che è congruente con $1 \pmod 4$. Tutti i suoi fattori primi sono maggiori di p_k; inoltre, dal punto (4) della dimostrazione precedente, non ha alcun fattore primo della forma $4m + 3$. Pertanto M_k ha un fattore primo della forma $4m + 1$ che è maggiore di p_k.

Due considerazioni chiudono la nostra discussione:

- Se a e b sono due numeri naturali primi fra loro, allora vi sono infiniti numeri primi della forma $am + b$ $(m \in \mathbb{N})$ — questo è un celebre (e difficile) teorema di Dirichlet. Più precisamente, si può dimostrare che il numero di primi $p \leq x$ della forma $p = am + b$ è descritto molto accuratamente per x grandi dalla funzione $\frac{1}{\varphi(a)} \frac{x}{\log x}$, dove $\varphi(a)$ indica il numero di b, tali che $1 \leq b < a$, che sono primi con a. (Questo è un sostanziale affinamento del teorema dei numeri primi, che abbiamo discusso a pagina 10.)
- Ciò significa che i numeri primi per a fissato e b variabile appaiono essenzialmente con la stessa frequenza. Ciononostante, ad esempio per $a = 4$, si può osservare un'alquanto sottile, ma al contempo percettibile e persistente tendenza verso più primi della forma $4m+3$: considerando un x grande arbitrario, vi sono buone probabilità che esistano più numeri primi $p \leq x$ della forma $p = 4m + 3$ che della forma $p = 4m + 1$. Questo effetto è noto come "Chebyshev's bias"; si vedano Riesel [4] e Rubinstein & Sarnak [5].

Bibliografia

[1] F. W. CLARKE, W. N. EVERITT, L. L. LITTLEJOHN & S. J. R. VORSTER: *H. J. S. Smith and the Fermat Two Squares Theorem,* Amer. Math. Monthly **106** (1999), 652-665.

[2] D. R. HEATH-BROWN: *Fermat's two squares theorem,* Invariant (1984), 2-5.

[3] I. NIVEN & H. S. ZUCKERMAN: *An Introduction to the Theory of Numbers,* Fifth edition, Wiley, New York 1972.

[4] H. RIESEL: *Prime Numbers and Computer Methods for Factorization,* Second edition, Progress in Mathematics **126**, Birkhäuser, Boston MA 1994.

[5] M. RUBINSTEIN & P. SARNAK: *Chebyshev's bias,* Experimental Mathematics **3** (1994), 173-197.

[6] A. THUE: *Et par antydninger til en taltheoretisk metode,* Kra. Vidensk. Selsk. Forh. **7** (1902), 57-75.

[7] S. WAGON: *Editor's corner: The Euclidean algorithm strikes again,* Amer. Math. Monthly **97** (1990), 125-129.

[8] D. ZAGIER: *A one-sentence proof that every prime* $p \equiv 1 \pmod 4$ *is a sum of two squares,* Amer. Math. Monthly **97** (1990), 144.

Ogni corpo finito è un campo

Capitolo 5

Gli anelli sono strutture importanti nell'algebra moderna. Se un anello R ha un elemento moltiplicativo unitario 1 ed ogni elemento non nullo ha un inverso moltiplicativo, allora R è chiamato *corpo*. Dunque, ciò che manca ad R per essere un campo è la commutatività della moltiplicazione. L'esempio più noto di corpo non commutativo è l'anello dei quaternioni scoperto da Hamilton. Ma, come dice il titolo del capitolo, un corpo non commutativo deve necessariamente essere infinito. Se R è finito, gli assiomi costringono la moltiplicazione ad essere commutativa.

Questo risultato, ormai un classico, ha attratto l'immaginazione di molti matematici, poiché, come scrisse Herstein: "Mette inaspettatamente in relazione due entità apparentemente non correlate come il numero di elementi in un certo sistema algebrico e la moltiplicazione di quel sistema".

Ernst Witt

Teorema. *Ogni corpo finito R è commutativo.*

Questo splendido teorema, generalmente attribuito a MacLagan Wedderburn, è stato dimostrato da diverse persone usando approcci assai diversi. Lo stesso Wedderburn fornì tre dimostrazioni nel 1905 ed un'altra dimostrazione fu data da Leonard E. Dickson nello stesso anno. Altre dimostrazioni furono date in seguito da Emil Artin, Hans Zassenhaus, Nicolas Bourbaki, e molti altri. C'è una dimostrazione che si distingue tuttavia per semplicità ed eleganza. Fu trovata da Ernst Witt nel 1931 e combina due idee elementari per estrarne una brillante conclusione.

■ **Dimostrazione.** Il nostro primo ingrediente deriva da una miscela di algebra lineare e teoria elementare dei gruppi. Per un elemento arbitrario $s \in R$, sia C_s l'insieme $\{x \in R : xs = sx\}$ degli elementi che commutano con s; C_s è chiamato il *centralizzatore* di s. Chiaramente, C_s contiene 0 e 1 ed è un sottocorpo di R. Il *centro* Z è l'insieme degli elementi che commutano con tutti gli elementi di R, pertanto $Z = \bigcap_{s \in R} C_s$. In particolare, tutti gli elementi di Z commutano, 0 e 1 appartengono a Z, e dunque Z è un *campo finito*. Poniamo $|Z| = q$.

Possiamo considerare R e C_s come spazi vettoriali sul campo Z e dedurre che $|R| = q^n$, dove n è la dimensione dello spazio vettoriale R su Z, ed analogamente $|C_s| = q^{n_s}$, per opportuni interi $n_s \geq 1$.

Supponiamo ora che R non sia un campo. Questo significa che per *un qualche* $s \in R$ il *centralizzatore* C_s non coincide con R, ovvero, che $n_s < n$.

Nell'insieme $R^* := R\backslash\{0\}$ consideriamo la relazione

$$r' \sim r \quad :\Longleftrightarrow \quad r' = x^{-1}rx \ \text{ per qualche } x \in R^*.$$

È semplice verificare che $\sim$ è una relazione di equivalenza. Sia

$$A_s \ := \ \{x^{-1}sx : x \in R^*\}$$

la classe di equivalenza cui appartiene s. Notiamo che $|A_s| = 1$ precisamente quando s è il centro di Z. Pertanto, secondo la nostra supposizione, vi sono classi A_s tali che $|A_s| \geq 2$. Si consideri ora per $s \in R^*$ la trasformazione $f_s : x \longmapsto x^{-1}sx$ di R^* su A_s. Per $x, y \in R^*$ troviamo

$$\begin{aligned} x^{-1}sx = y^{-1}sy \quad &\Longleftrightarrow \quad (yx^{-1})s = s(yx^{-1}) \\ &\Longleftrightarrow \quad yx^{-1} \in C_s^* \ \Longleftrightarrow \ y \in C_s^* x, \end{aligned}$$

dove $C_s^* := C_s\backslash\{0\}$, e $C_s^* x = \{zx : z \in C_s^*\}$ ha dimensione $|C_s^*|$. Dunque ogni elemento $x^{-1}sx$ è l'immagine di esattamente $|C_s^*| = q^{n_s} - 1$ elementi in R^* secondo la trasformazione f_s, e deduciamo $|R^*| = |A_s|\,|C_s^*|$. In particolare, notiamo che

$$\frac{|R^*|}{|C_s^*|} \ = \ \frac{q^n - 1}{q^{n_s} - 1} \ = \ |A_s| \quad \text{è un numero } \textit{intero} \text{ per ogni } s.$$

Sappiamo che le classi di equivalenza individuano una partizione su R^*. Raggruppiamo ora gli elementi centrali Z^* e designamo con $A_1, \ldots, A_t$ le classi di equivalenza contenenti più di un elemento. Grazie alla nostra supposizione sappiamo che $t \geq 1$. Poiché $|R^*| = |Z^*| + \sum_{k=1}^{t} |A_k|$, abbiamo dimostrato la cosiddetta *formula di classe*

$$q^n - 1 \ = \ q - 1 + \sum_{k=1}^{t} \frac{q^n - 1}{q^{n_k} - 1}, \tag{1}$$

dove per ogni k abbiamo $1 < \frac{q^n-1}{q^{n_k}-1} \in \mathbb{N}$.
Con (1) abbiamo lasciato l'algebra astratta e siamo tornati ai numeri naturali. Ora affermiamo che $q^{n_k} - 1 \,|\, q^n - 1$ implica $n_k \,|\, n$. In effetti, si scriva $n = an_k + r$ con $0 \leq r < n_k$: allora $q^{n_k} - 1 \,|\, q^{an_k+r} - 1$ implica

$$q^{n_k} - 1 \,|\, (q^{an_k+r} - 1) - (q^{n_k} - 1) = q^{n_k}(q^{(a-1)n_k+r} - 1),$$

e pertanto $q^{n_k} - 1 \,|\, q^{(a-1)n_k+r} - 1$, poiché q^{n_k} e $q^{n_k} - 1$ sono primi tra loro. Continuando in questo modo troviamo $q^{n_k} - 1 \,|\, q^r - 1$ con $0 \leq r < n_k$, il che è possibile solo per $r = 0$, ovvero, $n_k \,|\, n$. Riassumendo,

$$n_k \,|\, n \quad \text{per ogni } k. \tag{2}$$

Ecco ora il secondo ingrediente: i numeri complessi $\mathbb{C}$. Consideriamo il polinomio $x^n - 1$. Le sue radici in $\mathbb{C}$ sono dette le *radici n-esime dell'unità*. Poiché $\lambda^n = 1$, tutte queste radici λ hanno modulo unitario, $|\lambda| = 1$, e si

trovano pertanto sul cerchio unitario del piano complesso. In effetti, si tratta precisamente dei numeri $\lambda_k = e^{\frac{2k\pi i}{n}} = \cos(2k\pi/n) + i\sin(2k\pi/n)$, $0 \le k \le n-1$ (si veda il riquadro sotto). Alcune delle radici λ soddisfano $\lambda^d = 1$ per $d < n$; per esempio, la radice $\lambda = -1$ soddisfa $\lambda^2 = 1$. Per una radice λ, sia d il più piccolo esponente positivo con $\lambda^d = 1$, ovvero, d è l'ordine di λ nel gruppo delle radici dell'unità. Allora $d \mid n$, grazie al teorema di Lagrange ("l'ordine di ogni elemento di un gruppo divide l'ordine del gruppo" — si veda il riquadro nel Capitolo 1). Si noti che vi sono radici di ordine n, come $\lambda_1 = e^{\frac{2\pi i}{n}}$.

Radici dell'unità

Ogni numero complesso $z = x + iy$ può essere scritto in forma "polare"

$$z = re^{i\varphi} = r(\cos\varphi + i\sin\varphi),$$

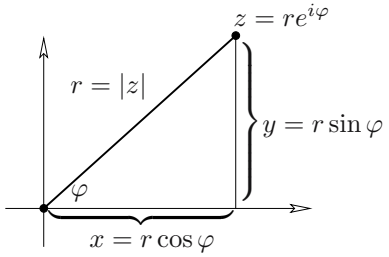
dove $r = |z| = \sqrt{x^2+y^2}$ è la distanza di z dall'origine e φ è l'angolo misurato a partire dal semiasse positivo x. Le radici n-esime dell'unità sono pertanto della forma

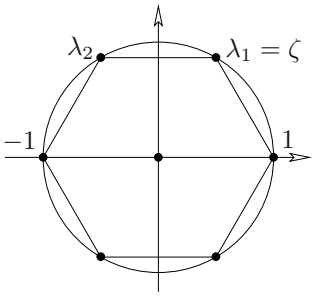
$$\lambda_k = e^{\frac{2k\pi i}{n}} = \cos(2k\pi/n) + i\sin(2k\pi/n), \qquad 0 \le k \le n-1,$$

poiché per ogni k

$$\lambda_k^n = e^{2k\pi i} = \cos(2k\pi) + i\sin(2k\pi) = 1.$$

Otteniamo queste radici geometricamente inscrivendo un poligono regolare a n lati nel cerchio unitario. Si noti che $\lambda_k = \zeta^k$ per ogni k, dove $\zeta = e^{\frac{2\pi i}{n}}$. Pertanto le radici n-esime dell'unità formano un gruppo ciclico $\{\zeta, \zeta^2, \ldots, \zeta^{n-1}, \zeta^n = 1\}$ di ordine n.





Le radici dell'unità per $n = 6$

Raggruppiamo ora tutte le radici di ordine d e poniamo

$$\phi_d(x) := \prod_{\lambda \text{ di ordine } d} (x - \lambda).$$

Si noti che la definizione di $\phi_d(x)$ è indipendente da n. Poiché ogni radice ha un certo ordine d, concludiamo che

$$x^n - 1 = \prod_{d \mid n} \phi_d(x). \tag{3}$$

Ecco l'osservazione cruciale: i *coefficienti* dei polinomi $\phi_n(x)$ sono *interi* (ovvero, $\phi_n(x) \in \mathbb{Z}[x]$ per ogni n) ed inoltre il coefficiente costante è o 1 o -1.

Verifichiamo con attenzione questa affermazione. Per $n = 1$ abbiamo 1 come unica radice, e dunque $\phi_1(x) = x - 1$. Ora procediamo per induzione,

supponendo $\phi_d(x) \in \mathbb{Z}[x]$ per ogni $d < n$, e che il coefficiente costante di $\phi_d(x)$ sia 1 o -1. Grazie a (3),

$$x^n - 1 \;=\; p(x)\,\phi_n(x) \tag{4}$$

dove $p(x) = \sum_{j=0}^{\ell} p_j x^j,\ \phi_n(x) = \sum_{k=0}^{n-\ell} a_k x^k,$ con $p_0 = 1$ o $p_0 = -1$.

Poiché $-1 = p_0 a_0$, vediamo che $a_0 \in \{1, -1\}$. Supponiamo di sapere già che $a_0, a_1, \ldots, a_{k-1} \in \mathbb{Z}$. Calcolando il coefficiente di x^k in entrambi i membri di (4) troviamo

$$\sum_{j=0}^{k} p_j a_{k-j} \;=\; \sum_{j=1}^{k} p_j a_{k-j} + p_0 a_k \;\in\; \mathbb{Z}.$$

Per ipotesi, tutti gli $a_0, \ldots, a_{k-1}$ (e tutti i p_j) sono in $\mathbb{Z}$. Pertanto $p_0 a_k$ e dunque a_k devono anch'essi essere interi, poiché p_0 vale 1 o -1.

Siamo pronti per il *coup de grâce* (NdT: in francese nell'edizione inglese). Sia $n_k \mid n$ uno dei numeri che compaiono in (1). Allora

$$x^n - 1 \;=\; \prod_{d \mid n} \phi_d(x) \;=\; (x^{n_k} - 1)\phi_n(x) \prod_{d \mid n,\, d \nmid n_k,\, d \neq n} \phi_d(x).$$

Concludiamo che in $\mathbb{Z}$ abbiamo le relazioni di divisibilità

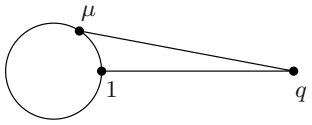
$$\phi_n(q) \mid q^n - 1 \qquad \text{e} \qquad \phi_n(q) \;\Big|\; \frac{q^n - 1}{q^{n_k} - 1}. \tag{5}$$

Poiché (5) vale per ogni k, deduciamo dalla formula (1)

$$\phi_n(q) \mid q - 1,$$

ma ciò non può essere vero. Perché? Sappiamo che $\phi_n(x) = \prod(x - \lambda)$ dove λ descrive tutte le radici di $x^n - 1$ di ordine n. Sia $\widetilde{\lambda} = a + ib$ una di queste radici. Dal momento che $n > 1$ (poiché $R \neq Z$) abbiamo $\widetilde{\lambda} \neq 1$, che implica che la parte reale a è più piccola di 1. Ora $|\widetilde{\lambda}|^2 = a^2 + b^2 = 1$, e pertanto

$$\begin{aligned} |q - \widetilde{\lambda}|^2 \;&=\; |q - a - ib|^2 \;=\; (q-a)^2 + b^2 \\ &=\; q^2 - 2aq + a^2 + b^2 \;=\; q^2 - 2aq + 1 \\ &>\; q^2 - 2q + 1 \qquad \text{(poiché } a < 1\text{)} \\ &=\; (q-1)^2, \end{aligned}$$



$|q - \mu| > |q - 1|$

e quindi $|q - \widetilde{\lambda}| > q - 1$ vale per *tutte* le radici di ordine n. Ciò implica

$$|\phi_n(q)| \;=\; \prod_{\lambda} |q - \lambda| > q - 1,$$

il che significa che $\phi_n(q)$ non può essere un divisore di $q - 1$. Questo porta ad una contraddizione e alla fine della dimostrazione. □

Bibliografia

[1] L. E. DICKSON: *On finite algebras,* Nachrichten der Akad. Wissenschaften Göttingen Math.-Phys. Klasse (1905), 1-36; Collected Mathematical Papers Vol. III, Chelsea Publ. Comp, The Bronx, NY 1975, 539-574.

[2] J. H. M. WEDDERBURN: *A theorem on finite algebras,* Trans. Amer. Math. Soc. **6** (1905), 349-352.

[3] E. WITT: *Über die Kommutativität endlicher Schiefkörper,* Abh. Math. Sem. Univ. Hamburg **8** (1931), 413.

Alcuni numeri irrazionali

Capitolo 6

> *"π è irrazionale"*

Questo risultato fu congetturato già da Aristotele, quando affermò che il diametro e la circonferenza di un cerchio non sono commensurabili. La prima dimostrazione di questo fatto fondamentale fu data da Johann Heinrich Lambert nel 1766. La nostra *Book Proof* risale al 1947 ed è dovuta a Ivan Niven: si tratta di una dimostrazione estremamente elegante di una pagina, che richiede soltanto l'analisi elementare. L'idea è notevole e se ne può derivare ben di più, come fu mostrato da Iwamoto e Koksma, rispettivamente:

- π^2 è irrazionale e
- e^r è irrazionale per $r \neq 0$ razionale.

Il metodo di Niven ha tuttavia radici e predecessori: si può infatti far risalire ad un lavoro classico di Charles Hermite del 1873, nel quale si stabilì per la prima volta che e è trascendente, ovvero che e non è lo zero di alcun polinomio con coefficienti razionali.

Charles Hermite

$$\begin{aligned} e &:= 1 + \tfrac{1}{1} + \tfrac{1}{2} + \tfrac{1}{6} + \tfrac{1}{24} + \dots \\ &= 2.718281828... \end{aligned}$$

Prima di considerare π analizzeremo e e le sue potenze, e vedremo che queste sono irrazionali. Questo è molto più semplice e per giunta seguiremo anche l'ordine storico nello sviluppo dei risultati. Per cominciare, è piuttosto semplice notare (come fece Fourier nel 1815) che $e = \sum_{k\geq 0} \frac{1}{k!}$ è irrazionale. In effetti, se vi fossero due interi a e $b > 0$ tali che $e = \frac{a}{b}$ avremmo allora

$$n!be \;=\; n!a$$

per *ogni* $n \geq 0$. Tuttavia ciò non può essere vero, poiché a destra abbiamo un intero, mentre il termine di sinistra, essendo

$$e \;=\; \Big(1+\frac{1}{1!}+\frac{1}{2!}+\ldots+\frac{1}{n!}\Big) + \Big(\frac{1}{(n+1)!}+\frac{1}{(n+2)!}+\frac{1}{(n+3)!}+\dots\Big),$$

si decompone in una parte intera

$$bn!\Big(1+\frac{1}{1!}+\frac{1}{2!}+\ldots+\frac{1}{n!}\Big)$$

ed in una seconda parte

$$b\Big(\frac{1}{n+1}+\frac{1}{(n+1)(n+2)}+\frac{1}{(n+1)(n+2)(n+3)}+\dots\Big)$$

Serie geometrica

Per la serie geometrica infinita

$$Q = \tfrac{1}{q} + \tfrac{1}{q^2} + \tfrac{1}{q^3} + \dots$$

con $q > 1$ abbiamo chiaramente

$$qQ = 1 + \tfrac{1}{q} + \tfrac{1}{q^2} + \dots = 1 + Q$$

e pertanto

$$Q = \frac{1}{q-1}.$$

che è *approssimativamente* $\frac{b}{n}$, cosicché per n grandi non può certamente essere intera: è maggiore di $\frac{b}{n+1}$e minore di $\frac{b}{n}$, come si può notare confrontandola con una serie geometrica:

$$\begin{aligned}\frac{1}{n+1} &< \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \frac{1}{(n+1)(n+2)(n+3)} + \dots \\ &< \frac{1}{n+1} + \frac{1}{(n+1)^2} + \frac{1}{(n+1)^3} + \dots = \frac{1}{n}.\end{aligned}$$

Qualcuno può essere indotto a pensare che questo semplice trucco di moltiplicare per $n!$ non sia neppure sufficiente per mostrare che e^2 è irrazionale. Questa è un'affermazione più forte: $\sqrt{2}$ è un esempio di numero irrazionale il cui quadrato non è irrazionale.

Da John Cosgrave abbiamo appreso che con due belle idee/osservazioni (chiamiamole trucchi) si può avanzare nondimeno di due passi: ognuno dei trucchi è sufficiente per dimostrare che e^2 è irrazionale, mentre la combinazione di entrambi consente persino di concludere che lo stesso vale per e^4. Il primo trucco si può trovare in una pubblicazione di una pagina ad opera di J. Liouville nel 1840 — mentre il secondo è in un "addendum" di due pagine che Liouville pubblicò nelle due pagine successive della stessa rivista.

Perché e^2 è irrazionale? Che cosa possiamo dedurre dall'uguaglianza $e^2 = \frac{a}{b}$? Seguendo Liouville dovremmo riscriverla come

$$be = ae^{-1},$$

sostituire le serie

$$e = 1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \dots$$

e

$$e^{-1} = 1 - \frac{1}{1} + \frac{1}{2} - \frac{1}{6} + \frac{1}{24} - \frac{1}{120} \pm \dots,$$

e quindi moltiplicare per $n!$, per un n pari sufficientemente grande. Quindi vediamo che $n!be$ è quasi intero e precisamente che

$$n!b\Big(1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{6} + \dots + \frac{1}{n!}\Big)$$

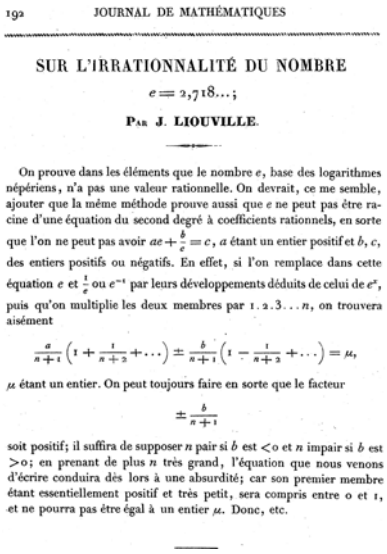
è un intero, ed il resto

$$n!b\Big(\frac{1}{(n+1)!} + \frac{1}{(n+2)!} + \dots\Big)$$

vale all'incirca $\frac{b}{n}$: è più grande di $\frac{b}{n+1}$ ma più piccolo di $\frac{b}{n}$, come abbiamo visto precedentemente.

Al contempo $n!ae^{-1}$ è anch'esso quasi intero: di nuovo abbiamo una grande parte intera ed un resto

$$(-1)^{n+1}n!a\Big(\frac{1}{(n+1)!} - \frac{1}{(n+2)!} + \frac{1}{(n+3)!} \mp \dots\Big),$$



192 JOURNAL DE MATHÉMATIQUES

SUR L'IRRATIONNALITÉ DU NOMBRE
$e = 2{,}718\ldots$;

PAR J. LIOUVILLE.

On prouve dans les éléments que le nombre e, base des logarithmes népériens, n'a pas une valeur rationnelle. On devrait, ce me semble, ajouter que la même méthode prouve aussi que e ne peut pas être racine d'une équation du second degré à coefficients rationnels, en sorte que l'on ne peut pas avoir $ae + \frac{b}{e} = c$, a étant un entier positif et b, c, des entiers positifs ou négatifs. En effet, si l'on remplace dans cette équation e et $\frac{1}{e}$ ou e^{-1} par leurs développements déduits de celui de e^x, puis qu'on multiplie les deux membres par $1.2.3\ldots n$, on trouvera aisément

$$\tfrac{a}{n+1}\Big(1 + \tfrac{1}{n+2} + \ldots\Big) \pm \tfrac{b}{n+1}\Big(1 - \tfrac{1}{n+2} + \ldots\Big) = \mu,$$

μ étant un entier. On peut toujours faire en sorte que le facteur

$$\pm \tfrac{b}{n+1}$$

soit positif; il suffira de supposer n pair si b est <0 et n impair si b est >0; en prenant de plus n très grand, l'équation que nous venons d'écrire conduira dès lors à une absurdité; car son premier membre étant essentiellement positif et très petit, sera compris entre 0 et 1, et ne pourra pas être égal à un entier μ. Donc, etc.

La pubblicazione di Liouville

e ciò vale all'incirca $(-1)^{n+1}\frac{a}{n}$. Più precisamente, per n pari il resto è maggiore di $-\frac{a}{n}$, ma minore di

$$-a\Big(\frac{1}{n+1}-\frac{1}{(n+1)^2}-\frac{1}{(n+1)^3}-\dots\Big) \;=\; -\frac{a}{n+1}\Big(1-\frac{1}{n}\Big) \;<\; 0.$$

Ma ciò non può essere vero, poiché per n grandi e pari implicherebbe che $n!ae^{-1}$ è solo leggermente inferiore ad un intero, mentre $n!be$ è leggermente superiore ad un intero, dunque $n!ae^{-1} = n!be$ non potrebbe valere. □

Per dimostrare che e^4 è irrazionale, supporremo ora con un po' di coraggio che $e^4 = \frac{a}{b}$ sia razionale e lo scriveremo come

$$be^2 \;=\; ae^{-2}.$$

Potremmo ora provare a moltiplicare per $n!$ per un n grande e raccogliere gli addendi non interi, ma ciò non conduce a nulla di utile: la somma dei termini rimanenti a sinistra varrà all'incirca $b\frac{2^{n+1}}{n}$, a destra $(-1)^{n+1}a\frac{2^{n+1}}{n}$, ed entrambi saranno molto grandi se n diventa grande.

Pertanto bisogna esaminare la situazione con un po' più di attenzione e operare due piccoli aggiustamenti alla strategia: innanzitutto non prenderemo un n grande *arbitrario*, ma una potenza grande di due, $n = 2^m$; in secondo luogo non moltiplicheremo per $n!$ ma per $\frac{n!}{2^{n-1}}$. Ora abbiamo bisogno di un piccolo lemma, un caso speciale del teorema di Legendre (si veda a pagina 9): per ogni $n \geq 1$ il numero intero $n!$ contiene il fattore primo 2 al più $n-1$ volte, e l'uguaglianza vale se (e solo se) n è una potenza di due, $n = 2^m$.

Questo lemma non è di difficile dimostrazione: $\lfloor\frac{n}{2}\rfloor$ dei fattori di $n!$ sono pari, $\lfloor\frac{n}{4}\rfloor$ di essi sono divisibili per 4, e così via. Quindi se 2^k è la più grande delle due potenze che soddisfi $2^k \leq n$, allora $n!$ contiene il fattore primo 2 esattamente

$$\Big\lfloor\frac{n}{2}\Big\rfloor+\Big\lfloor\frac{n}{4}\Big\rfloor+\ldots+\Big\lfloor\frac{n}{2^k}\Big\rfloor \;\leq\; \frac{n}{2}+\frac{n}{4}+\ldots+\frac{n}{2^k} \;=\; n\Big(1-\frac{1}{2^k}\Big) \;\leq\; n-1$$

volte, con uguaglianza in entrambe le disequazioni se $n = 2^k$.

Torniamo ora a $be^2 = ae^{-2}$. Siamo in cerca di

$$b\frac{n!}{2^{n-1}}e^2 \;=\; a\frac{n!}{2^{n-1}}e^{-2} \tag{1}$$

e sostituiamo la serie

$$e^2 \;=\; 1+\frac{2}{1}+\frac{4}{2}+\frac{8}{6}+\;\ldots\;+\frac{2^r}{r!}+\ldots$$

e

$$e^{-2} \;=\; 1-\frac{2}{1}+\frac{4}{2}-\frac{8}{6}\pm\;\ldots\;+(-1)^r\frac{2^r}{r!}+\ldots$$

Per $r \leq n$ otteniamo addendi interi da entrambe le parti, ovvero

$$b\frac{n!}{2^{n-1}}\frac{2^r}{r!} \quad \text{risp.} \quad (-1)^r a\frac{n!}{2^{n-1}}\frac{2^r}{r!},$$

dove per $r > 0$ il denominatore $r!$ contiene il fattore primo 2 *al più* $r-1$ volte, mentre $n!$ lo contiene *esattamente* $n-1$ volte (dunque per $r > 0$ gli addendi sono pari).

Poiché n è pari (avendo supposto che $n = 2^m$), le serie che otteniamo per $r \geq n+1$ sono

$$2b\Big(\frac{2}{n+1} + \frac{4}{(n+1)(n+2)} + \frac{8}{(n+1)(n+2)(n+3)} + \ \ldots\Big)$$

risp.

$$2a\Big(-\frac{2}{n+1} + \frac{4}{(n+1)(n+2)} - \frac{8}{(n+1)(n+2)(n+3)} \pm \ \ldots\Big).$$

Per n grandi queste serie varranno all'incirca $\frac{4b}{n}$ risp. $-\frac{4a}{n}$, come si può vedere ancora per analogia con le serie geometriche. Per $n = 2^m$ grandi questo implica che il membro di sinistra di (1) è *un po'* più grande di un intero, mentre il membro di destra è *un po'* più piccolo, il che è una contraddizione! □

Sappiamo dunque che e^4 è irrazionale; per mostrare che e^3, e^5 etc. sono anch'essi irrazionali, abbiamo bisogno di un macchinario più potente (cioè di un po' di analisi), e di una nuova idea — che essenzialmente rimanda a Charles Hermite, e la cui chiave è nascosta nel semplice lemma seguente.

Lemma. *Per un $n \geq 1$ fissato, sia*

$$f(x) \ = \ \frac{x^n(1-x)^n}{n!}.$$

(i) *La funzione $f(x)$ è un polinomio della forma $f(x) = \dfrac{1}{n!}\displaystyle\sum_{i=n}^{2n} c_i x^i$, dove i coefficienti c_i sono numeri interi.*

(ii) *Per $0 < x < 1$ abbiamo $0 < f(x) < \frac{1}{n!}$.*

(iii) *Le derivate $f^{(k)}(0)$ e $f^{(k)}(1)$ sono numeri interi per ogni $k \geq 0$.*

■ **Dimostrazione.** Le affermazioni (i) e (ii) sono evidenti.
Per dimostrare (iii) si noti che da (i) la k-esima derivata $f^{(k)}$ si annulla in $x = 0$ a meno che $n \leq k \leq 2n$, ed in questo intervallo $f^{(k)}(0) = \frac{k!}{n!}c_k$ è intero. Da $f(x) = f(1-x)$ otteniamo $f^{(k)}(x) = (-1)^k f^{(k)}(1-x)$ per ogni x, e pertanto $f^{(k)}(1) = (-1)^k f^{(k)}(0)$, che è un intero. □

Teorema 1. *e^r è irrazionale per ogni $r \in \mathbb{Q}\backslash\{0\}$.*

■ **Dimostrazione.** È sufficiente mostrare che e^s non può essere razionale per un numero intero positivo s (se $e^{\frac{s}{t}}$ fosse razionale, allora anche $\big(e^{\frac{s}{t}}\big)^t = e^s$ sarebbe razionale). Supponiamo che $e^s = \frac{a}{b}$ per due interi $a, b > 0$, e n sia tanto grande che $n! > as^{2n+1}$. Si ponga

La stima $n! > e(\frac{n}{e})^n$ dà esplicitamente un n "abbastanza grande"

$$F(x) \ := \ s^{2n}f(x) - s^{2n-1}f'(x) + s^{2n-2}f''(x) \mp \ \ldots \ + f^{(2n)}(x),$$

dove $f(x)$ è la funzione del lemma.

$F(x)$ può anche essere scritta come la somma infinita

$$F(x) \;=\; s^{2n}f(x) - s^{2n-1}f'(x) + s^{2n-2}f''(x) \mp \;\ldots,$$

poiché le derivate di ordine più alto $f^{(k)}(x)$, per $k > 2n$, si annullano. Da questo fatto possiamo vedere che il polinomio $F(x)$ soddisfa l'identità

$$F'(x) \;=\; -s\,F(x) + s^{2n+1}f(x).$$

Pertanto,

$$\frac{d}{dx}\left[e^{sx}F(x)\right] \;=\; se^{sx}F(x) + e^{sx}F'(x) \;=\; s^{2n+1}e^{sx}f(x)$$

e dunque

$$N \;:=\; b\int_0^1 s^{2n+1}e^{sx}f(x)dx \;=\; b\left[e^{sx}F(x)\right]_0^1 \;=\; aF(1) - bF(0).$$

Questo è un intero, poiché la parte (iii) del lemma implica che $F(0)$ e $F(1)$ sono interi. Peraltro, la parte (ii) del lemma fornisce delle stime superiori e inferiori per la dimensione di N,

$$0 \;<\; N \;=\; b\int_0^1 s^{2n+1}e^{sx}f(x)dx \;<\; bs^{2n+1}e^s\frac{1}{n!} \;=\; \frac{as^{2n+1}}{n!} \;<\; 1,$$

il che mostra che N non può essere un numero intero: da qui la contraddizione. □

Visto che questo trucco si è rivelato così efficace, lo usiamo ancora una volta.

Teorema 2. *π^2 è irrazionale.*

■ **Dimostrazione.** Supponiamo che $\pi^2 = \frac{a}{b}$ per due interi $a, b > 0$. Usiamo ora il polinomio

$$F(x) \;:=\; b^n\Big(\pi^{2n}f(x) - \pi^{2n-2}f^{(2)}(x) + \pi^{2n-4}f^{(4)}(x) \mp \ldots\Big),$$

che soddisfa $F''(x) = -\pi^2F(x) + b^n\pi^{2n+2}f(x)$.
Dalla parte (iii) del lemma otteniamo che $F(0)$ e $F(1)$ sono numeri interi. Regole elementari di derivazione danno

$$\begin{aligned}\frac{d}{dx}\big[F'(x)\sin\pi x - \pi F(x)\cos\pi x\big] \;&=\; \big(F''(x) + \pi^2F(x)\big)\sin\pi x\\ &=\; b^n\pi^{2n+2}f(x)\sin\pi x\\ &=\; \pi^2a^nf(x)\sin\pi x,\end{aligned}$$

e perciò otteniamo

$$\begin{aligned}N := \pi\int_0^1 a^nf(x)\sin\pi x\,dx \;&=\; \Big[\frac{1}{\pi}F'(x)\sin\pi x - F(x)\cos\pi x\Big]_0^1\\ &=\; F(0) + F(1),\end{aligned}$$

π non è razionale, ma può essere "ben approssimato" da razionali – alcuni di questi approssimanti sono noti sin dall'antichità:

$$\begin{aligned}\tfrac{22}{7} \;&=\; 3.142857142857...\\ \tfrac{355}{113} \;&=\; 3.141592920353...\\ \tfrac{104348}{33215} \;&=\; 3.141592653921...\\ \pi \;&=\; 3.141592653589...\end{aligned}$$

che è un intero. Inoltre N è positivo poiché è definito come l'integrale di una funzione positiva (tranne sul bordo). Tuttavia, se scegliamo un n così grande che $\frac{\pi a^n}{n!} < 1$, allora dalla parte (ii) del lemma otteniamo

$$0 \;<\; N \;=\; \pi \int_0^1 a^n f(x) \sin \pi x \, dx \;<\; \frac{\pi a^n}{n!} \;<\; 1,$$

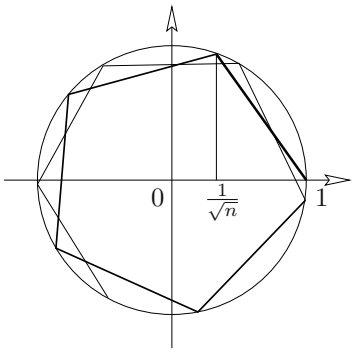
e quindi una contraddizione. □

Ecco dunque il nostro risultato finale sull'irrazionalità.

Teorema 3. *Per ogni intero dispari $n \geq 3$, il numero*

$$A(n) \;:=\; \frac{1}{\pi} \arccos\left(\frac{1}{\sqrt{n}}\right)$$

è irrazionale.



Avremo bisogno di questo risultato per il terzo problema di Hilbert (si veda il Capitolo 8) nei casi $n = 3$ e $n = 9$. Per $n = 2$ e $n = 4$ abbiamo $A(2) = \frac{1}{4}$ e $A(4) = \frac{1}{3}$, quindi la restrizione agli interi dispari è essenziale. Questi valori si derivano facilmente facendo appello al diagramma a margine, in cui l'affermazione "$\frac{1}{\pi} \arccos\left(\frac{1}{\sqrt{n}}\right)$ è irrazionale" è equivalente a dire che l'arco poligonale costruito da $\frac{1}{\sqrt{n}}$, le cui corde hanno tutte la stessa lunghezza, non si chiude mai in se stesso.

Lasciamo al lettore come esercizio la dimostrazione che $A(n)$ è razionale *solo* per $n \in \{1, 2, 4\}$. Per provarlo, si distinguano i casi in cui $n = 2^r$ e in cui n non sia una potenza di 2.

■ **Dimostrazione.** Usiamo il teorema dell'addizione

$$\cos\alpha + \cos\beta = 2\cos\tfrac{\alpha+\beta}{2} \cos\tfrac{\alpha-\beta}{2}$$

dalla trigonometria elementare, che per $\alpha = (k+1)\varphi$ e $\beta = (k-1)\varphi$ dà

$$\cos(k+1)\varphi \;=\; 2\cos\varphi \;\cos\; k\varphi \;-\; \cos(k-1)\varphi. \tag{2}$$

Per l'angolo $\varphi_n = \arccos\left(\frac{1}{\sqrt{n}}\right)$, definito da $\cos\varphi_n = \frac{1}{\sqrt{n}}$ e $0 \leq \varphi_n \leq \pi$, si ottengono rappresentazioni della forma

$$\cos\; k\varphi_n \;=\; \frac{A_k}{\sqrt{n}^k},$$

dove A_k è un intero non divisibile per n, per ogni $k \geq 0$. In effetti, abbiamo una tale rappresentazione per $k = 0, 1$ con $A_0 = A_1 = 1$, e per induzione su k usando (2) otteniamo per $k \geq 1$

$$\cos(k+1)\varphi_n \;=\; 2\frac{1}{\sqrt{n}}\frac{A_k}{\sqrt{n}^k} - \frac{A_{k-1}}{\sqrt{n}^{k-1}} \;=\; \frac{2A_k - nA_{k-1}}{\sqrt{n}^{k+1}}.$$

Abbiamo pertanto $A_{k+1} = 2A_k - nA_{k-1}$. Se $n \geq 3$ è dispari e A_k non è divisibile per n, allora troviamo che A_{k+1} non può essere neppure divisibile per n.

Supponiamo ora che

$$A(n) \;=\; \frac{1}{\pi}\varphi_n \;=\; \frac{k}{\ell}$$

sia razionale (con interi $k, \ell > 0$). Allora $\ell\varphi_n = k\pi$ dà

$$\pm 1 \;=\; \cos\, k\pi \;=\; \frac{A_\ell}{\sqrt{n}^{\ell}}.$$

Pertanto $\sqrt{n}^{\ell} \;=\; \pm A_\ell$ è un intero, con $\ell \geq 2$, e dunque $n \,|\, \sqrt{n}^{\ell}$. Con $\sqrt{n}^{\ell} \,|\, A_\ell$ troviamo che n divide A_ℓ, il che è una contraddizione. □

Bibliografia

[1] C. HERMITE: *Sur la fonction exponentielle,* Comptes rendus de l'Académie des Sciences (Paris) **77** (1873), 18-24; Œuvres de Charles Hermite, Vol. III, Gauthier-Villars, Paris 1912, pp. 150-181.

[2] Y. IWAMOTO: *A proof that π^2 is irrational,* J. Osaka Institute of Science and Technology **1** (1949), 147-148.

[3] J. F. KOKSMA: *On Niven's proof that π is irrational,* Nieuw Archief voor Wiskunde (2) **23** (1949), 39.

[4] J. LIOUVILLE: *Sur l'irrationalité du nombre $e = 2{,}718...$,* Journal de Mathématiques Pures et Appl. (1) **5** (1840), 192; *Addition*, 193-194.

[5] I. NIVEN: *A simple proof that π is irrational,* Bulletin Amer. Math. Soc. **53** (1947), 509.

Tre volte $\pi^2/6$

Capitolo 7

Sappiamo che la serie infinita $\sum_{n\geq 1}\frac{1}{n}$ non converge. In effetti, nel Capitolo 1 abbiamo visto che persino la serie $\sum_{p\in\mathbb{P}}\frac{1}{p}$ diverge.
Tuttavia, la somma dei reciproci dei quadrati converge (seppur molto lentamente, come vedremo) e produce un valore interessante.

Serie di Eulero.

$$\sum_{n\geq 1}\frac{1}{n^2} \;=\; \frac{\pi^2}{6}.$$

Questo è un classico, celebre ed importante risultato di Eulero del 1734. Una delle interpretazioni chiave di questo risultato è che esso dà il primo valore non banale $\zeta(2)$ della funzione zeta di Riemann (si veda l'appendice di pagina 47). Tale valore è irrazionale, come abbiamo visto nel Capitolo 6.
Ma non solo questo risultato ha un ruolo prominente nella storia della matematica: esiste anche un buon numero di dimostrazioni estremamente eleganti ed intelligenti che hanno una storia propria. Per alcune di queste dimostrazioni la gioia della scoperta e della riscoperta è stata condivisa da molti. In questo capitolo ne presentiamo tre.

1	$=$	1.000000
$1+\frac{1}{4}$	$=$	1.250000
$1+\frac{1}{4}+\frac{1}{9}$	$=$	1.361111
$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{6}$	$=$	1.423611
$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{6}+\frac{1}{25}$	$=$	1.463611
$1+\frac{1}{4}+\frac{1}{9}+\frac{1}{6}+\frac{1}{25}+\frac{1}{36}$	$=$	1.491388
$\pi^2/6$	$=$	1.644934

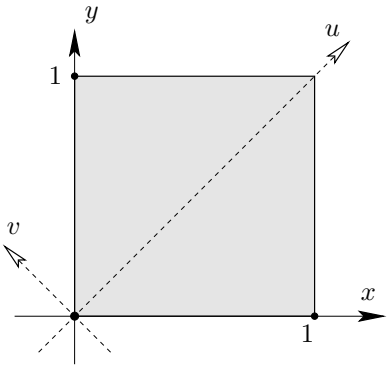
■ **Dimostrazione.** La prima dimostrazione appare sotto forma di esercizio nel libro sulla teoria dei numeri di William J. LeVeque nel 1956. Ma egli dice: "Non ho la minima idea di quale fosse l'origine di quel problema, ma sono piuttosto certo che non fosse mia".
La dimostrazione consiste in due diverse valutazioni dell'integrale doppio

$$I \;:=\; \int_0^1\!\!\int_0^1 \frac{1}{1-xy}\,dx\,dy.$$

Per il primo, sviluppiamo $\frac{1}{1-xy}$ in una serie geometrica, scomponiamo gli addendi sotto forma di prodotti ed integriamo senza fatica:

$$\begin{aligned} I &= \int_0^1\int_0^1 \sum_{n\ge0}(xy)^n\,dx\,dy \;=\; \sum_{n\ge0}\int_0^1\int_0^1 x^n y^n\,dx\,dy \\ &= \sum_{n\ge0}\Big(\int_0^1 x^n dx\Big)\Big(\int_0^1 y^n dy\Big) \;=\; \sum_{n\ge0}\frac{1}{n+1}\,\frac{1}{n+1} \\ &= \sum_{n\ge0}\frac{1}{(n+1)^2} \;=\; \sum_{n\ge1}\frac{1}{n^2} \;=\; \zeta(2). \end{aligned}$$

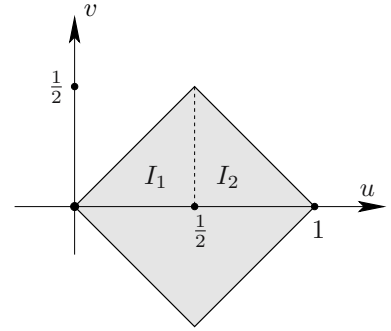
Questo calcolo mostra anche che l'integrale doppio (di una funzione positiva con un polo in $x=y=1$) è finito. Si noti che il calcolo è semplice e diretto anche se lo leggiamo al contrario — pertanto la stima di $\zeta(2)$ porta all'integrale doppio I.



Il secondo modo di valutare I viene da un cambio di coordinate: nelle nuove coordinate date da $u := \frac{y+x}{2}$ e $v := \frac{y-x}{2}$ il dominio di integrazione è un quadrato di lato $\frac12\sqrt2$, che otteniamo dal vecchio dominio, prima ruotandolo di $45°$ e poi riducendolo di un fattore $\sqrt2$. Le sostituzioni $x = u - v$ e $y = u + v$ danno

$$\frac{1}{1-xy} \;=\; \frac{1}{1-u^2+v^2}.$$

Per trasformare l'integrale, dobbiamo sostituire $dx\,dy$ con $2\,du\,dv$, per compensare il fatto che la nostra trasformazione di coordinate riduce le aree di un fattore costante di 2 (che è il determinante jacobiano della trasformazione; si veda il riquadro alla pagina successiva). Il nuovo dominio d'integrazione, e la funzione da integrare, sono simmetrici rispetto all'asse u, pertanto dobbiamo soltanto calcolare due volte (e qui esce un altro fattore 2) l'integrale sulla metà superiore del dominio, che dividiamo in due parti nel modo più naturale:



$$I \;=\; 4\int_0^{1/2}\Big(\int_0^u \frac{dv}{1-u^2+v^2}\Big)du \;+\; 4\int_{1/2}^1\Big(\int_0^{1-u}\frac{dv}{1-u^2+v^2}\Big)du.$$

Usando $\displaystyle\int \frac{dx}{a^2+x^2} = \frac1a\arctan\frac{x}{a} + C$, ciò diventa

$$\begin{aligned} I \;=\;& 4\int_0^{1/2}\frac{1}{\sqrt{1-u^2}}\arctan\Big(\frac{u}{\sqrt{1-u^2}}\Big)\,du \\ &+\; 4\int_{1/2}^1\frac{1}{\sqrt{1-u^2}}\arctan\Big(\frac{1-u}{\sqrt{1-u^2}}\Big)\,du. \end{aligned}$$

Questi integrali possono essere semplificati ed infine calcolati sostituendo $u = \sin\theta$, risp. $u = \cos\theta$. Ma procediamo più direttamente osservando che la derivata di $g(u) := \arctan\big(\frac{u}{\sqrt{1-u^2}}\big)$ è $g'(u) = \frac{1}{\sqrt{1-u^2}}$, mentre quella

di $h(u) := \arctan\left(\frac{1-u}{\sqrt{1-u^2}}\right) = \arctan\left(\sqrt{\frac{1-u}{1+u}}\right)$ è $h'(u) = -\frac{1}{2}\frac{1}{\sqrt{1-u^2}}$. Possiamo dunque usare $\int_a^b f'(x)f(x)dx = \left[\frac{1}{2}f(x)^2\right]_a^b = \frac{1}{2}f(b)^2 - \frac{1}{2}f(a)^2$ ottenendo

$$\begin{aligned} I &= 4\int_0^{1/2} g'(u)g(u)\,du \;+\; 4\int_{1/2}^1 -2h'(u)h(u)\,du \\ &= 2\Big[g(u)^2\Big]_0^{1/2} \;-\; 4\Big[h(u)^2\Big]_{1/2}^1 \\ &= 2g(\tfrac{1}{2})^2 \;-\; 2g(0)^2 \;-\; 4h(1)^2 \;+\; 4h(\tfrac{1}{2})^2 \\ &= 2\big(\tfrac{\pi}{6}\big)^2 - 0 - 0 + 4\big(\tfrac{\pi}{6}\big)^2 \;=\; \tfrac{\pi^2}{6}. \end{aligned}$$ □

Questa dimostrazione fornisce la somma della serie di Eulero come valore di un integrale effettuando una semplice trasformazione di coordinate. Una dimostrazione ingegnosa di questo tipo — con una trasformazione di coordinate del tutto non banale — fu scoperta in seguito da Beukers, Calabi e Kolk. Il punto di partenza è scomporre la somma $\sum_{n\ge1}\frac{1}{n^2}$ nei suoi termini pari e dispari. Chiaramente la somma dei termini pari $\frac{1}{2^2} + \frac{1}{4^2} + \frac{1}{6^2} + \ldots = \sum_{k\ge1}\frac{1}{(2k)^2}$ dà $\frac{1}{4}\zeta(2)$, dunque i termini dispari $\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \ldots = \sum_{k\ge0}\frac{1}{(2k+1)^2}$ danno i tre quarti della somma totale $\zeta(2)$. Pertanto la serie di Eulero equivale a

$$\sum_{k\ge0}\frac{1}{(2k+1)^2} \;=\; \frac{\pi^2}{8}.$$

■ **Dimostrazione.** Come sopra, possiamo esprimere questa somma come un integrale doppio, ovvero

$$J \;=\; \int_0^1\int_0^1 \frac{1}{1-x^2y^2}\,dx\,dy \;=\; \sum_{k\ge0}\frac{1}{(2k+1)^2}.$$

Per calcolare J, Beukers, Calabi e Kolk proposero le nuove coordinate

$$u \;:=\; \arccos\sqrt{\frac{1-x^2}{1-x^2y^2}} \qquad v \;:=\; \arccos\sqrt{\frac{1-y^2}{1-x^2y^2}}.$$

Per calcolare l'integrale doppio, possiamo ignorare il bordo del dominio e considerare x, y nell'intervallo $0 < x < 1$ e $0 < y < 1$. Dunque u, v si troverà nel triangolo $u > 0$, $v > 0$, $u + v < \pi/2$. La trasformazione di coordinate può essere invertita esplicitamente fornendo

$$x \;=\; \frac{\sin u}{\cos v} \qquad \text{e} \qquad y \;=\; \frac{\sin v}{\cos u}.$$

La formula di sostituzione

Per calcolare un integrale doppio

$$I \;=\; \int_S f(x,y)\,dx\,dy$$

possiamo operare una sostituzione di variabili

$$x = x(u,v) \quad y = y(u,v),$$

se la corrispondenza fra $(u,v) \in T$ e $(x,y) \in S$ è biiettiva e differenziabile con continuità. Quindi I è uguale a

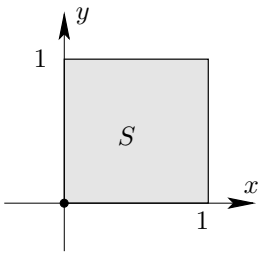
$$\int_T f(x(u,v),y(u,v))\left|\frac{d(x,y)}{d(u,v)}\right|du\,dv,$$

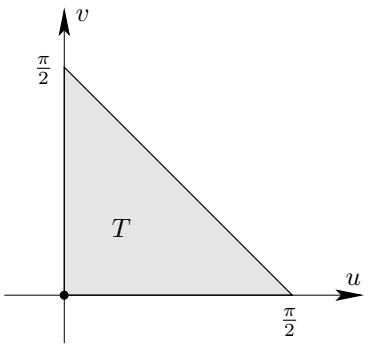
dove $\frac{d(x,y)}{d(u,v)}$ è il determinante jacobiano:

$$\frac{d(x,y)}{d(u,v)} \;=\; \det\begin{pmatrix} \frac{dx}{du} & \frac{dx}{dv} \\ \frac{dy}{du} & \frac{dy}{dv} \end{pmatrix}.$$

È facile verificare che queste formule definiscono una trasformazione di coordinate biiettiva tra l'interno del quadrato unitario $S = \{(x, y) : 0 \le x, y \le 1\}$ e l'interno del triangolo $T = \{(u, v) : u, v \ge 0,\ u + v \le \pi/2\}$. Dobbiamo ora calcolare il determinante jacobiano della trasformazione di coordinate, che magicamente si rivela essere

$$\det \begin{pmatrix} \frac{\cos u}{\cos v} & \frac{\sin u \sin v}{\cos^2 v} \\ \frac{\sin u \sin v}{\cos^2 u} & \frac{\cos v}{\cos u} \end{pmatrix} \;=\; 1 - \frac{\sin^2 u \sin^2 v}{\cos^2 u \cos^2 v} \;=\; 1 - x^2 y^2.$$





Per $m = 1, 2, 3$ la formula a fianco fornisce
$\cot^2 \frac{\pi}{3} = \frac{1}{3}$
$\cot^2 \frac{\pi}{5} + \cot^2 \frac{2\pi}{5} = 2$
$\cot^2 \frac{\pi}{7} + \cot^2 \frac{2\pi}{7} + \cot^2 \frac{3\pi}{7} = 5$

Ma ciò significa che l'integrale che vogliamo calcolare si trasforma in

$$J \;=\; \int_0^{\pi/2} \int_0^{\pi/2-u} 1 \, du \, dv,$$

che è proprio l'area $\frac{1}{2}(\frac{\pi}{2})^2 = \frac{\pi^2}{8}$ del triangolo T. □

Splendido — tanto più che lo stesso metodo di dimostrazione si estende al calcolo di $\zeta(2k)$ come un integrale $2k$-dimensionale, per ogni $k \ge 1$. Ci riferiamo alla pubblicazione originale di Beuker, Calabi e Kolk [2], ed al Capitolo 20, dove otterremo questo risultato attraverso un percorso diverso, tramite il trucco di Herglotz e l'approccio originale di Eulero.

Dopo queste due dimostrazioni basate su trasformazioni di coordinate, non possiamo resistere alla tentazione di presentare un'altra dimostrazione, completamente diversa e del tutto elementare, del risultato $\sum_{n\ge1} \frac{1}{n^2} = \frac{\pi^2}{6}$. Essa appare in una serie di esercizi in un libro di problemi dei gemelli Akiva e Isaak Yaglom, la cui edizione russa originale apparve nel 1954. Versioni di questa splendida dimostrazione furono riscoperte e presentate da F. Holme (1970), I. Papadimitriou (1973), e da Ransford (1982) che la attribuì a John Scholes.

■ **Dimostrazione.** Il primo passo è stabilire una relazione notevole tra i valori della funzione cotangente (al quadrato). Precisamente, per ogni $m \ge 1$ si ha

$$\cot^2\left(\tfrac{\pi}{2m+1}\right) + \cot^2\left(\tfrac{2\pi}{2m+1}\right) + \ldots + \cot^2\left(\tfrac{m\pi}{2m+1}\right) \;=\; \tfrac{2m(2m-1)}{6}. \quad (1)$$

Per dimostrarlo, cominciamo con la relazione

$$\cos nx + i \sin nx \;=\; (\cos x + i \sin x)^n$$

e prendiamone la parte immaginaria, che è

$$\sin nx \;=\; \binom{n}{1} \sin x \cos^{n-1} x - \binom{n}{3} \sin^3 x \cos^{n-3} x \pm \ldots \qquad (2)$$

Ora poniamo $n = 2m + 1$, mentre per x considereremo gli m valori diversi $x = \frac{r\pi}{2m+1}$, per $r = 1, 2, \ldots, m$. Per ognuno di questi valori abbiamo $nx = r\pi$, e pertanto $\sin nx = 0$, mentre $0 < x < \frac{\pi}{2}$ implica che per $\sin x$ otteniamo m valori positivi distinti.

In particolare, possiamo dividere (2) per $\sin^n x$, che dà

$$0 \;=\; \binom{n}{1}\cot^{n-1}x - \binom{n}{3}\cot^{n-3}x \pm \ldots,$$

ovvero,

$$0 \;=\; \binom{2m+1}{1}\cot^{2m}x - \binom{2m+1}{3}\cot^{2m-2}x \pm \ldots$$

per ognuno degli m valori distinti di x. Pertanto per il polinomio di grado m

$$p(t) \;:=\; \binom{2m+1}{1}t^m - \binom{2m+1}{3}t^{m-1} \pm \ldots + (-1)^m\binom{2m+1}{2m+1}$$

conosciamo m radici distinte

$$a_r \;=\; \cot^2\left(\tfrac{r\pi}{2m+1}\right) \quad \text{for} \quad r = 1, 2, \ldots, m.$$

Dunque il polinomio coincide con

$$p(t) \;=\; \binom{2m+1}{1}\left(t - \cot^2\left(\tfrac{\pi}{2m+1}\right)\right)\cdot\ldots\cdot\left(t - \cot^2\left(\tfrac{m\pi}{2m+1}\right)\right).$$

Confronto di coefficienti:
se $p(t) = c(t-a_1)\cdots(t-a_m)$,
allora il coefficiente di t^{m-1} è
$-c(a_1 + \ldots + a_m)$

Confrontando i coefficienti di t^{m-1} in $p(t)$ si ottiene ora che la somma delle radici è

$$a_1 + \ldots + a_r \;=\; \frac{\binom{2m+1}{3}}{\binom{2m+1}{1}} \;=\; \tfrac{2m(2m-1)}{6},$$

il che dimostra (1).

Ci serve anche una seconda identità dello stesso tipo,

$$\csc^2\left(\tfrac{\pi}{2m+1}\right) + \csc^2\left(\tfrac{2\pi}{2m+1}\right) + \ldots + \csc^2\left(\tfrac{m\pi}{2m+1}\right) = \tfrac{2m(2m+2)}{6}, \qquad (3)$$

per la funzione cosecante $\csc x = \frac{1}{\sin x}$. Ma

$$\csc^2 x = \frac{1}{\sin^2 x} = \frac{\cos^2 x + \sin^2 x}{\sin^2 x} = \cot^2 x + 1,$$

cosicché possiamo derivare (3) da (1) aggiungendo m ad entrambi i termini dell'equazione.

Ora il contesto è definito e tutto prende il proprio posto. Sfruttiamo il fatto che nell'intervallo $0 < y < \frac{\pi}{2}$ abbiamo

$$0 \;<\; \sin y \;<\; y \;<\; \tan y,$$

$0 < a < b < c$
implica
$0 < \frac{1}{c} < \frac{1}{b} < \frac{1}{a}$

e pertanto

$$0 \;<\; \cot y \;<\; \tfrac{1}{y} \;<\; \csc y,$$

che implica

$$\cot^2 y \;<\; \tfrac{1}{y^2} \;<\; \csc^2 y.$$

Ora consideriamo questa doppia disuguaglianza, applichiamola ad ognuno degli m valori distinti di x, e sommiamo i risultati. Usando (1) per il termine di sinistra e (3) per quello di destra, otteniamo

$$\tfrac{2m(2m-1)}{6} \;<\; \left(\tfrac{2m+1}{\pi}\right)^2 + \left(\tfrac{2m+1}{2\pi}\right)^2 + \ldots + \left(\tfrac{2m+1}{m\pi}\right)^2 \;<\; \tfrac{2m(2m+2)}{6},$$

ovvero,

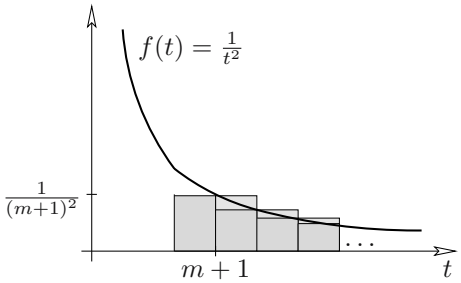
$$\tfrac{\pi^2}{6}\tfrac{2m}{2m+1}\tfrac{2m-1}{2m+1} \;<\; \tfrac{1}{1^2} + \tfrac{1}{2^2} + \ldots + \tfrac{1}{m^2} \;<\; \tfrac{\pi^2}{6}\tfrac{2m}{2m+1}\tfrac{2m+2}{2m+1}.$$

Entrambi i termini di destra e di sinistra convergono a $\frac{\pi^2}{6}$ for $m \longrightarrow \infty$: fine della dimostrazione. □

Ora, quanto velocemente $\sum \frac{1}{n^2}$ converge a $\pi^2/6$? Per determinarlo dobbiamo stimare la differenza

$$\frac{\pi^2}{6} - \sum_{n=1}^{m} \frac{1}{n^2} \;=\; \sum_{n=m+1}^{\infty} \frac{1}{n^2}.$$

Questo è molto semplice con la tecnica del "confronto con un integrale" che abbiamo già considerato nell'appendice al Capitolo 2 (pagina 10). Essa dà



$$\sum_{n=m+1}^{\infty} \frac{1}{n^2} \;<\; \int_m^{\infty} \frac{1}{t^2}dt \;=\; \frac{1}{m}$$

come stima dall'alto e

$$\sum_{n=m+1}^{\infty} \frac{1}{n^2} \;>\; \int_{m+1}^{\infty} \frac{1}{t^2}dt \;=\; \frac{1}{m+1}$$

come stima dal basso sugli "addendi rimasti" — o addirittura

$$\sum_{n=m+1}^{\infty} \frac{1}{n^2} \;>\; \int_{m+\frac{1}{2}}^{\infty} \frac{1}{t^2}dt \;=\; \frac{1}{m+\frac{1}{2}}$$

se desiderate operare una stima leggermente più attenta, sfruttando il fatto che la funzione $f(t) = \frac{1}{t^2}$ è convessa.

Ciò significa che la nostra serie non converge troppo bene; se sommiamo i primi mille addendi, ci aspettiamo un errore sulla terza cifra dopo la virgola, mentre per la somma del primo milione di addendi, $m = 1000000$, ci aspettiamo un errore sulla sesta cifra decimale, che infatti otteniamo. Tuttavia, ecco una grande sorpresa: con una precisione di 45 cifre,

$$\begin{aligned} \pi^2/6 &= 1.64493406684822643647241516664602518921894990 1, \\ \sum_{n=1}^{10^6} \frac{1}{n^2} &= 1.64493306684872643630574849997939185588561654 4. \end{aligned}$$

Pertanto la sesta cifra dopo la virgola è errata (inferiore di 1), ma *le sei cifre successive sono giuste*! Quindi troviamo ancora una cifra sbagliata

(superiore di 5) ed altre cinque cifre corrette. Questa sorprendente scoperta è alquanto recente e si deve a Roy D. North di Colorado Springs, 1988. (Nel 1982, Martin R. Powell, un insegnante di Amersham, Bucks, in Inghilterra, non riuscì a notare l'effetto intero a causa dell'insufficiente potenza di calcolo disponibile all'epoca). È qualcosa di troppo strano per essere una pura coincidenza ... Uno sguardo al termine di errore, che di nuovo limitatamente alle prime 45 cifre dà

$$\sum_{n=10^6+1}^{\infty} \frac{1}{n^2} \;=\; 0.000000999999500000166666666666633333333333357,$$

rivela l'esistenza di una "trama". Si potrebbe provare a riscrivere quest'ultimo numero come

$$+\,10^{-6} \;-\; \tfrac{1}{2}10^{-12} \;+\; \tfrac{1}{6}10^{-18} \;-\; \tfrac{1}{30}10^{-30} \;+\; \tfrac{1}{42}10^{-42} \;+\; \ldots$$

dove i coefficienti $(1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42})$ di 10^{-6i} formano l'inizio della successione dei *numeri di Bernoulli* che ritroveremo nel Capitolo 20. Consigliamo ai nostri lettori l'articolo di Borwein, Borwein & Dilcher [3] per scoprire altre sorprendenti "coincidenze" — e per le dimostrazioni.

Appendice: la funzione zeta di Riemann

La *funzione zeta di Riemann* $\zeta(s)$ è definita per numeri reali $s > 1$ da

$$\zeta(s) \;:=\; \sum_{n\geq 1} \frac{1}{n^s}.$$

Le nostre stime per H_n (si veda a pagina 10) implicano che la serie diverge per $\zeta(1)$, ma converge per ogni $s > 1$ reale. La funzione zeta ha un prolungamento canonico all'intero piano complesso (con un polo semplice in $s = 1$), che può essere costruito mediante sviluppi in serie di potenze. La funzione complessa risultante è di essenziale importanza per la teoria dei numeri primi. Citiamo tre diverse connessioni:

(1) L'identità notevole

$$\zeta(s) \;=\; \prod_p \frac{1}{1-p^{-s}}$$

si deve ad Eulero. Essa codifica il fatto basilare che ogni numero naturale ha un'unica (!) scomposizione in fattori primi; grazie a questo risultato, l'identità di Eulero è una semplice conseguenza dello sviluppo della serie geometrica

$$\frac{1}{1-p^{-s}} \;=\; 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \ldots$$

(2) La posizione degli zeri complessi della funzione zeta è l'argomento dell'"ipotesi di Riemann": una delle più importanti ed irrisolte congetture

dell'intera matematica. Essa afferma che tutti gli zeri non banali $s \in \mathbb{C}$ della funzione zeta soddisfano $\mathrm{Re}(s) = \frac{1}{2}$. La funzione zeta si annulla in tutti gli interi pari negativi, che vengono detti "zeri banali".

Molto recentemente, Jeff Lagarias ha mostrato che, sorprendentemente, l'ipotesi di Riemann è equivalente alla seguente proprietà elementare: per ogni $n \geq 1$,

$$\sum_{d \,|\, n} d \;\leq\; H_n \,+\, \exp(H_n)\log(H_n),$$

con uguaglianza solo per $n = 1$, dove H_n è ancora l'n-esimo numero armonico.

(3) È noto da molto tempo che $\zeta(s)$ è un multiplo razionale di π^s, e pertanto irrazionale, se s è un intero *pari* $s \geq 2$; si veda il Capitolo 20. Al contrario, l'irrazionalità di $\zeta(3)$ fu dimostrata da Roger Apéry solamente nel 1979. Nonostante gli sforzi profusi non è ancora noto quanto valga $\zeta(s)$ per gli altri interi dispari, $s = 2t + 1 \geq 5$. Molto recentemente, Keith Ball e Tanguy Rivoal hanno dimostrato che un numero infinito dei valori $\zeta(2t+1)$ è irrazionale. In effetti, sebbene non si sappia se $\zeta(s)$ sia irrazionale per ogni singolo valore dispari $s \geq 5$, Wadim Zudilin ha dimostrato che almeno uno dei quattro valori $\zeta(5)$, $\zeta(7)$, $\zeta(9)$, e $\zeta(11)$ è irrazionale. Rinviamo il lettore alla splendida rassegna di Fischler [4].

Bibliografia

[1] K. BALL & T. RIVOAL: *Irrationalité d'une infinité de valeurs de la fonction zêta aux entiers impairs,* Inventiones math. **146** (2001), 193-207.

[2] F. BEUKERS, J. A. C. KOLK & E. CALABI: *Sums of generalized harmonic series and volumes,* Nieuw Archief voor Wiskunde (4) **11** (1993), 217-224.

[3] J. M. BORWEIN, P. B. BORWEIN & K. DILCHER: *Pi, Euler numbers, and asymptotic expansions,* Amer. Math. Monthly **96** (1989), 681-687.

[4] S. FISCHLER: *Irrationalité de valeurs de zêta (d'après Apéry, Rivoal, ...),* Bourbaki Seminar, No. 910, November 2002; in Astérisque; Preprint `arXiv:math.NT/0303066`, March 2003, 45 pages.

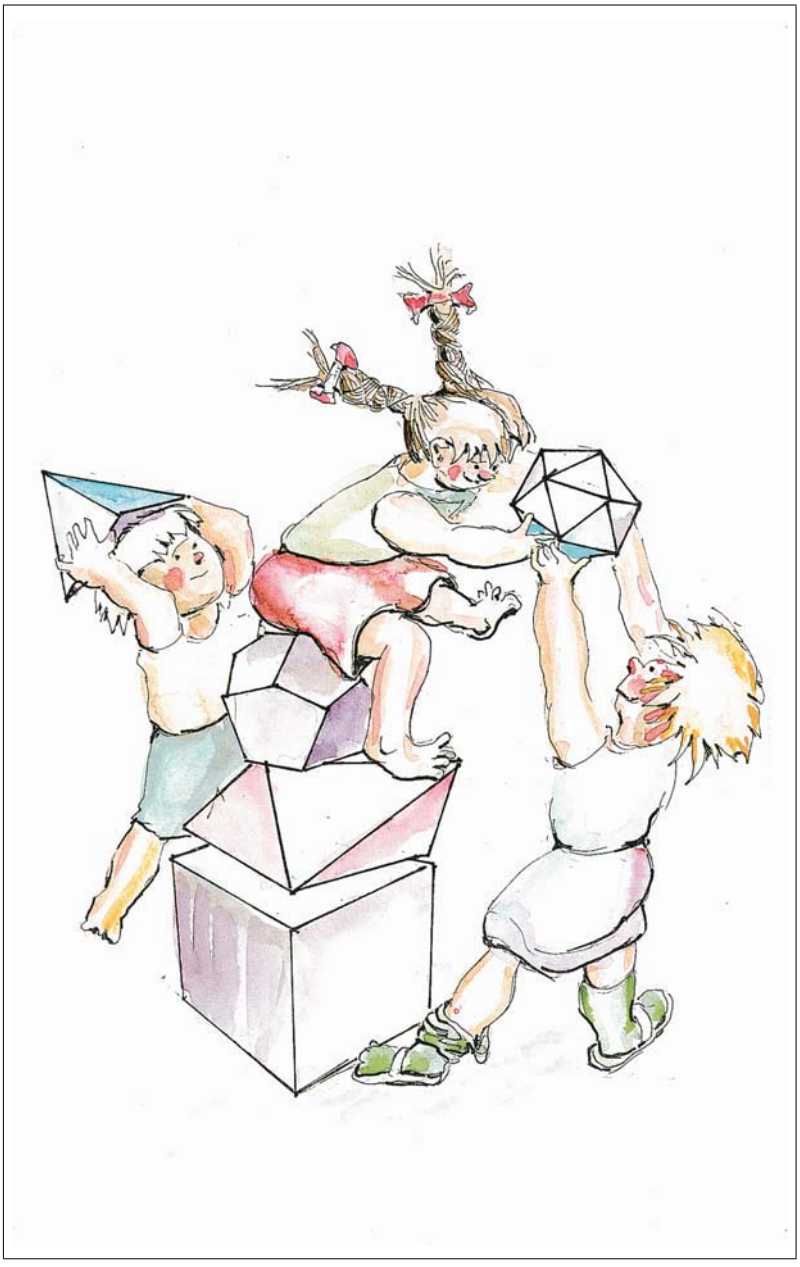
[5] J. C. LAGARIAS: *An elementary problem equivalent to the Riemann hypothesis,* Amer. Math. Monthly **109** (2002), 534-543.

[6] W. J. LEVEQUE: *Topics in Number Theory,* Vol. I, Addison-Wesley, Reading MA 1956.

[7] A. M. YAGLOM & I. M. YAGLOM: *Challenging mathematical problems with elementary solutions,* Vol. II, Holden-Day, Inc., San Francisco, CA 1967.

[8] W. ZUDILIN: *Arithmetic of linear forms involving odd zeta values,* Preprint, August 2001, 42 pages; `arXiv:math.NT/0206176`.

Geometria



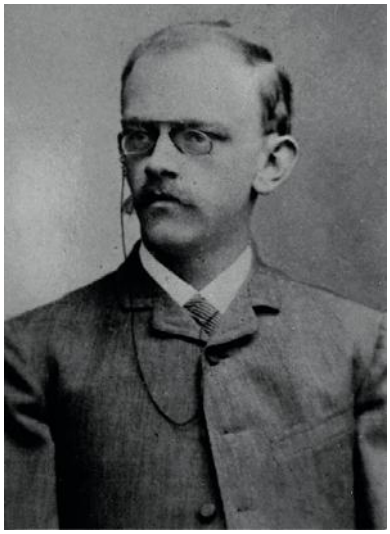
"Solidi platonici — un gioco da ragazzi!"

Il terzo problema di Hilbert: la scomposizione di poliedri

Capitolo 8

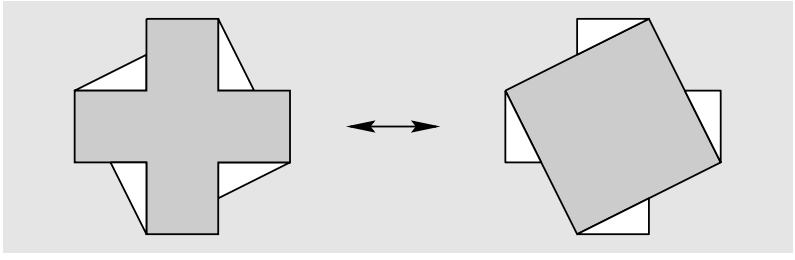
Nel suo leggendario intervento al Congresso Internazionale dei Matematici di Parigi nel 1900 David Hilbert chiese — come terzo dei suoi ventitré problemi — di specificare

> *"due tetraedri di basi uguali ed altezze uguali che non possano in alcun modo essere divisi in tetraedri congruenti, e che non possano essere combinati con tetraedri congruenti per formare due poliedri che possano a loro volta essere suddivisi in tetraedri congruenti"* .

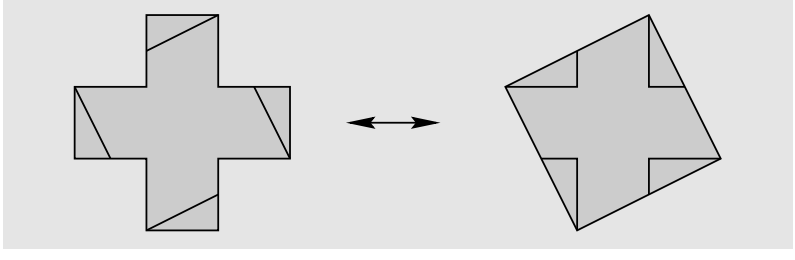


David Hilbert

Questo problema si può far risalire a due lettere di Carl Friedrich Gauss del 1844 (pubblicate nella raccolta dei lavori di Gauss nel 1900). Se dei tetraedri di uguale volume potessero essere suddivisi in parti congruenti, allora questo fornirebbe una dimostrazione "elementare" del teorema XII.5 di Euclide, che afferma che piramidi con stessa base e stessa altezza hanno ugual volume. Fornirebbe perciò una definizione elementare del volume per i poliedri (non derivante dall'analisi e dunque non basata su argomenti di continuità). Un'affermazione analoga è vera nella geometria piana: il Teorema di Bolyai-Gerwien [1, Sez. 2.7] afferma che i poligoni piani sono sia *equiscomponibili* (possono essere scomposti in triangoli congruenti) sia *equicomplementabili* (possono essere resi congruenti tramite l'aggiunta di triangoli congruenti) se e solo se hanno la stessa area.



La croce è equicomplementabile con un quadrato della stessa area



In effetti, sono persino equiscomponibili

Hilbert — come possiamo vedere dalla sua formulazione del problema — si aspettava che non vi fosse un teorema analogo in dimensione tre, ed aveva ragione. In effetti, il problema fu completamente risolto da uno studente di Hilbert, Max Dehn, in due pubblicazioni: la prima, che riporta tetraedri non equiscomponibili di base ed altezza uguali, apparve già nel 1900. La seconda, anch'essa relativa all'equicomplementabilità, apparve nel 1902. Tuttavia, le pubblicazioni di Dehn non sono di facile comprensione ed è difficile verificare se Dehn non sia caduto in una sottile trappola che colse altri: una molto-elegante-ma-putroppo-falsa dimostrazione fu trovata da Bricard (nel 1896!), da Meschkowski (1960) e probabilmente da altri. Fortunatamente, la dimostrazione di Dehn fu rimaneggiata e rifatta, e dopo gli sforzi combinati di V. F. Kagan (1903/1930), Hugo Hadwiger (1949/54) e Vladimir G. Boltianskii, abbiamo ora una *Book Proof* — come riportato di seguito. L'appendice di questo capitolo fornisce alcune nozioni basilari sui poliedri.

(1) Un po' di algebra lineare

Per ogni insieme finito di numeri reali $M = \{m_1, \ldots, m_k\} \subseteq \mathbb{R}$, definiamo $V(M)$ come l'insieme di tutte le combinazioni lineari dei numeri in M con coefficienti razionali, ovvero

$$V(M) := \Big\{ \sum_{i=1}^{k} q_i m_i : q_i \in \mathbb{Q} \Big\} \subseteq \mathbb{R}.$$

La prima osservazione (ovvia, ma importante) è che $V(M)$ è uno spazio vettoriale di dimensione finita sul campo dei numeri razionali $\mathbb{Q}$. In effetti, $V(M)$ è chiaramente chiuso rispetto alle somme e alle moltiplicazioni con i numeri razionali, e gli assiomi di campo per $\mathbb{R}$ rendono $V(M)$ uno spazio vettoriale. La dimensione di $V(M)$ è la cardinalità di un qualunque insieme generatore minimale. Poiché per definizione M genera $V(M)$, vediamo che esso contiene un sistema minimale di generatori, e pertanto

$$\dim_{\mathbb{Q}} V(M) \leq k = |M|.$$

In seguito dovremo usare delle *funzioni $\mathbb{Q}$-lineari*

$$f : V(M) \to \mathbb{Q}$$

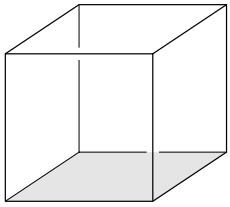
che interpretiamo come trasformazioni lineari di spazi $\mathbb{Q}$-vettoriali. La proprietà chiave è che per ogni relazione di dipendenza lineare $\sum_{i=1}^{k} q_i m_i = 0$ con $q_i \in \mathbb{Q}$, dobbiamo avere $\sum_{i=1}^{k} q_i f(m_i) = f(0) = 0$. Ecco il semplice lemma che fa funzionare le cose.

Lemma. *Per ogni sottoinsieme finito $M \subseteq M'$ di $\mathbb{R}$, lo spazio $\mathbb{Q}$-vettoriale $V(M)$ è un sottospazio dello spazio $\mathbb{Q}$-vettoriale $V(M')$. Pertanto se $f : V(M) \to \mathbb{Q}$ è una funzione $\mathbb{Q}$-lineare, allora f si può estendere ad una funzione $\mathbb{Q}$-lineare $f' : V(M') \to \mathbb{Q}$ così che $f'(m) = f(m)$ per ogni $m \in M$.*

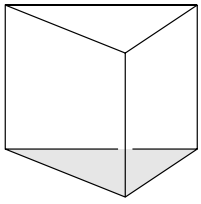
■ **Dimostrazione.** Ogni funzione $\mathbb{Q}$-lineare $V(M) \to \mathbb{Q}$ è determinata non appena siano fissati i suoi valori su una $\mathbb{Q}$-base di $V(M)$. Poiché ogni base di $V(M)$ si può estendere ad una base di $V(M')$, ne segue il resto. □

(2) Gli invarianti di Dehn

Per un poliedro tridimensionale P, si indichi con M_P l'insieme di tutti gli angoli tra facce adiacenti (*angoli diedri*), insieme al numero π. Pertanto per un cubo C otteniamo $M_C = \left\{\frac{\pi}{2}, \pi\right\}$, mentre per un prisma ortogonale Q su un triangolo equilatero otteniamo $M_Q = \left\{\frac{\pi}{3}, \frac{\pi}{2}, \pi\right\}$.



$M_C = \{\frac{\pi}{2}, \pi\}$



$M_Q = \{\frac{\pi}{3}, \frac{\pi}{2}, \pi\}$

Dato un qualsiasi insieme finito $M \subseteq \mathbb{R}$ che contenga M_P ed una qualsiasi funzione $\mathbb{Q}$-lineare

$$f : V(M) \to \mathbb{Q}$$

che soddisfi $f(\pi) = 0$, definiamo l'*invariante di Dehn* di P (rispetto a f) come il numero reale

$$D_f(P) \ := \ \sum_{e \in P} \ell(e) \cdot f(\alpha(e)),$$

dove la somma si estende a tutti gli spigoli e del poliedro, $\ell(e)$ indica la lunghezza di e ed $\alpha(e)$ è l'angolo tra le due facce che si incontrano in e.

In seguito calcoleremo diversi invarianti di Dehn. Per ora si noti soltanto che $f\left(\frac{\pi}{2}\right) = \frac{1}{2} f(\pi) = 0$ deve valere per *qualunque* funzione $\mathbb{Q}$-lineare f siffatta, e pertanto

$$D_f(C) \ = \ 0,$$

ovvero l'invariante di Dehn di un cubo è zero per *qualunque* f.

(3) Il teorema di Dehn-Hadwiger

Come sopra, chiamiamo due poliedri P, Q *equiscomponibili* se essi si possono scomporre in insiemi finiti di poliedri $P_1, \ldots, P_n$ e $Q_1, \ldots, Q_n$ tali che P_i e Q_i siano congruenti per ogni i $(1 \leq i \leq n)$. Due poliedri sono *equicomplementabili* se esistono dei poliedri $P_1, \ldots, P_m$ e $Q_1, \ldots, Q_m$ tali che gli interni dei P_i siano disgiunti fra loro e da P e lo stesso vale per i Q_i e Q, cosicché P_i è congruente a Q_i per ogni i, e $\widetilde{P} := P \cup P_1 \cup P_2 \cup \ldots \cup P_m$ e $\widetilde{Q} := Q \cup Q_1 \cup Q_2 \cup \ \ldots \ \cup Q_m$ sono equiscomponibili.

Un teorema di Gerling del 1844 implica che è irrilevante che si ammettano o meno le riflessioni nel considerare le congruenze.

Chiaramente poliedri equiscomponibili sono equicomplementabili, ma il reciproco è lungi dall'essere evidente. Il seguente teorema di Hadwiger (nella versione di Boltianskii) ci fornisce lo strumento per trovare — come propose Hilbert — tetraedri di uguale volume che non siano equicomplementabili e pertanto non equiscomponibili.

Teorema. *Siano P e Q dei poliedri con angoli diedri $\alpha_1, \ldots, \alpha_p$ rispettivamente $\beta_1, \ldots, \beta_q$, ai loro spigoli e sia M un insieme finito di numeri reali con*

$$\{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, \pi\} \subseteq M.$$

Se $f : V(M) \to \mathbb{Q}$ è una qualsiasi funzione $\mathbb{Q}$-lineare con $f(\pi) = 0$ tale che

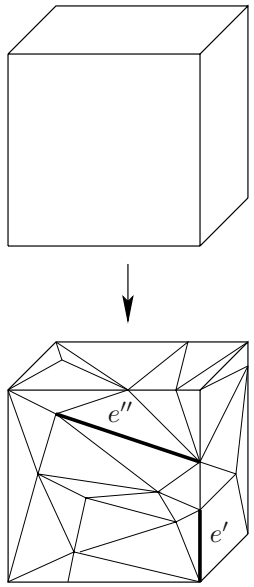
$$D_f(P) \neq D_f(Q),$$

allora P e Q non sono equicomplementabili.

■ **Dimostrazione.** L'argomentazione si articola in due parti.

(1) Se un poliedro P si scompone in un numero finito di poliedri $P_1, \ldots, P_n$, e se tutti gli angoli diedri delle singole parti $P_1, \ldots, P_n$ sono contenuti nell'insieme M, allora per ogni funzione $\mathbb{Q}$-lineare $f : V(M) \to \mathbb{Q}$, gli invarianti di Dehn si sommano:

$$D_f(P) \;=\; D_f(P_1) + \ldots + D_f(P_n).$$



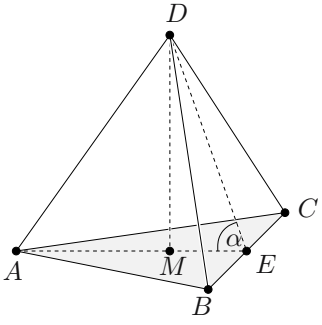
Per questo associamo una *massa* ad una qualsiasi parte di uno spigolo di un poliedro: se $e' \subseteq e$ è una parte di uno spigolo e di P, allora la sua massa sarà

$$m_f(e') \;:=\; \ell(e')\; f(\alpha(e')),$$

ovvero la sua lunghezza per l'f-valore del suo angolo diedro.

Ora se P si scompone in $P_1, \ldots, P_n$, si consideri l'unione di tutti gli spigoli delle parti P_i. Lungo gli spigoli e' contenuti in spigoli di P, vediamo che gli angoli diedri delle parti si sommano nell'angolo diedro di P in e' e pertanto le masse si sommano.

Ad un qualsiasi altro spigolo e'' di uno dei P_i contenuto nell'interno di una faccia di P o nell'interno di P, gli angoli si sommano dando π o 2π, così che gli f-valori degli angoli nelle parti si sommano dando $f(\pi) = 0$, rispettivamente $f(2\pi) = 0$. Pertanto per la somma delle masse otteniamo lo stesso valore che avevamo inizialmente associato a questi spigoli per P, ovvero 0.



(2) Supponendo che P e Q siano equicomplementabili, possiamo allargare M ad un soprainsieme M' che includa anche tutti gli angoli diedri che appaiono in ognuna delle parti considerate. M' è finito, poiché consideriamo solo scomposizioni finite. Quindi il lemma citato sopra permette di estendere f ad $f' : V(M') \to \mathbb{Q}$, e pertanto la parte **(1)** dà un'equazione del tipo

$$D_{f'}(P)+D_{f'}(P_1)+\ldots+D_{f'}(P_m) = D_{f'}(Q)+D_{f'}(Q_1)+\ldots+D_{f'}(Q_m)$$

dove $D_{f'}(P_i) = D_{f'}(Q_i)$ essendo P_i e Q_i congruenti. Quindi concludiamo che $D_f(P) = D_f(Q)$, il che è una contraddizione. □

Esempio 1. Per un tetraedro regolare T_0 con spigoli di lunghezza ℓ, calcoliamo l'angolo diedro servendoci della figura a margine. Il punto medio M del triangolo di base divide l'altezza AE dello stesso triangolo per 1:2, e poiché $|AE| = |DE|$, troviamo $\cos\alpha = \frac{1}{3}$ e dunque

$$\alpha \;=\; \arccos\tfrac{1}{3}.$$

Ponendo $M := \{\alpha, \pi\}$ notiamo che il rapporto

$$\frac{\alpha}{\pi} \;=\; \frac{1}{\pi}\arccos\tfrac{1}{3}$$

è irrazionale, secondo il Teorema 3 del Capitolo 6 (prendendo $n = 9$). Pertanto lo spazio $\mathbb{Q}$-vettoriale $V(M)$ è di dimensione due con base M e vi è una funzione $\mathbb{Q}$-lineare $f : V(M) \to \mathbb{Q}$ con

$$f(\alpha) := 1,\; f(\pi) := 0.$$

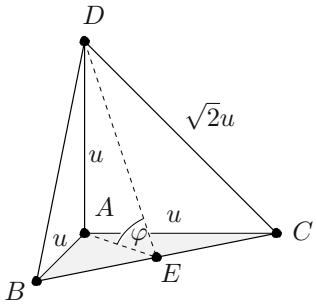
Per questa f abbiamo

$$D_f(T_0) \;=\; 6\ell f(\alpha) \;=\; 6\ell \;\neq\; 0,$$

di conseguenza un *tetraedro regolare non può essere equiscomponibile o equicomplementabile con un cubo*, dal momento che l'invariante di Dehn di un cubo si annulla per ogni f.

Esempio 2. Sia T_1un tetraedro individuato da tre spigoli ortogonali AB, AC, AD di lunghezza u. Tale tetraedro ha tre angoli diedri retti e tre altri angoli diedri di uguale ampiezza φ, che calcoliamo dalla traccia come

$$\cos\varphi \;=\; \frac{|AE|}{|DE|} \;=\; \frac{\frac{1}{2}\sqrt{2}u}{\frac{1}{2}\sqrt{3}\,\sqrt{2}u} \;=\; \frac{1}{\sqrt{3}}.$$



Ne segue che

$$\varphi \;=\; \arccos\frac{1}{\sqrt{3}}.$$

Per $M := \left\{\frac{\pi}{2}, \arccos\frac{1}{\sqrt{3}}, \pi\right\}$, lo spazio $\mathbb{Q}$-vettoriale $V(M)$ ha dimensione 2. In effetti, π e $\frac{\pi}{2}$ sono linearmente dipendenti, così $V(M) = V\left(\left\{\arccos\frac{1}{\sqrt{3}}, \pi\right\}\right)$, ma non vi è alcuna relazione razionale tra $\arccos\frac{1}{\sqrt{3}}$ e π — equivalentemente, $\frac{1}{\pi}\arccos\frac{1}{\sqrt{3}}$ è irrazionale, come abbiamo dimostrato nel Capitolo 6 (si prenda $n = 3$ nel Teorema 3).

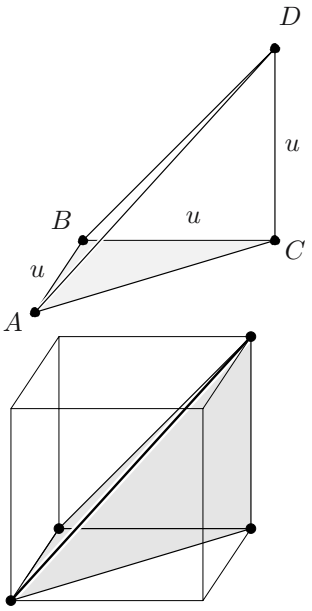
Pertanto possiamo costruire una trasformazione $\mathbb{Q}$-lineare f ponendo

$$f(\pi) := 0 \quad \text{e} \quad f\left(\arccos\tfrac{1}{\sqrt{3}}\right) := 1,$$

da cui otteniamo $f(\frac{\pi}{2}) = 0$ e quindi

$$D_f(T_1) \;=\; 3uf\left(\tfrac{\pi}{2}\right) + 3\left(\sqrt{2}u\right)f\left(\arccos\tfrac{1}{\sqrt{3}}\right) \;=\; 3\sqrt{2}u \;\neq\; 0.$$

Abbiamo allora dimostrato che T_1 non è equiscomponibile o equicomplementabile con un cubo C dello stesso volume, poiché $D_f(C) = 0$ vale per *qualunque* f.



Esempio 3. Infine, sia T_2 un tetraedro con tre spigoli consecutivi AB, BC e CD ortogonali tra loro (un "ortoschema") e di uguale lunghezza u.

Non calcoleremo gli angoli in questo tetraedro (essi sono $\frac{\pi}{2}$, $\frac{\pi}{3}$, e $\frac{\pi}{4}$), ma affermeremo piuttosto che — usando i punti medi degli spigoli e delle facce ed il centro — un cubo con spigoli di lunghezza u si può scomporre in 6 tetraedri di questo tipo (3 copie congruenti e 3 immagini speculari).

Tutte queste copie congruenti ed immagini speculari hanno gli stessi invarianti di Dehn e pertanto per ogni f funzionale appropriato otterremo

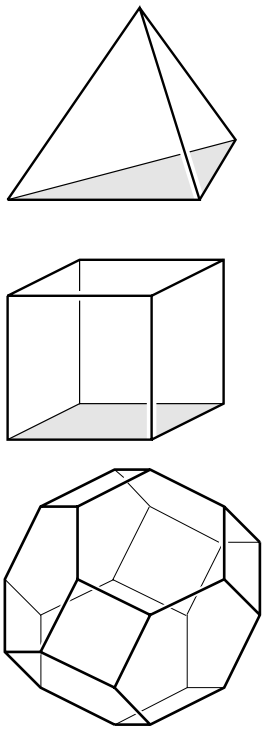
$$D_f(T_2) = \frac{1}{6} D_f(C) = 0$$

quindi tutti gli invarianti di Dehn di tale tetraedro si annullano! Questo risolve il terzo problema di Hilbert, poiché abbiamo costruito prima un tetraedro diverso, T_1, con basi congruenti e la stessa altezza e con $D_f(T_1) \neq 0$. Secondo il teorema di Dehn-Hadwiger T_1 e T_2 non sono equiscomponibili e nemmeno equicomplementabili.

Appendice: politopi e poliedri

Un *politopo convesso* in $\mathbb{R}^d$ è l'inviluppo convesso di un insieme finito $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, ovvero un insieme della forma

$$P \;=\; \text{conv}(S) \;:=\; \Big\{ \sum_{i=1}^{n} \lambda_i \boldsymbol{s}_i : \lambda_i \geq 0,\; \sum_{i=1}^{n} \lambda_i = 1 \Big\}.$$



Alcuni politopi familiari: tetraedro, cubo e permutaedro

I politopi sono certamente oggetti familiari: i principali esempi sono i *poligoni* convessi (politopi convessi bidimensionali) ed i *poliedri* convessi (politopi convessi tridimensionali).

Esistono diversi tipi di poliedri che si generalizzano a dimensioni più elevate in modo naturale. Per esempio, se l'insieme S di cardinalità $d+1$ non è contenuto in un sottospazio affine di dimensione inferiore a d, allora $\text{conv}(S)$ è un *simplesso* d-dimensionale (o *d-simplesso*). Per $d = 2$ abbiamo un triangolo, per $d = 3$ un tetraedro. Analogamente, quadrati e cubi sono casi speciali di d-cubi, come il *d-cubo unitario* dato da $C_d = [0,1]^d \subseteq \mathbb{R}^d$.

I politopi generali sono definiti come unioni finite di politopi convessi. In questo libro i poliedri non convessi appariranno in relazione con il teorema della rigidità di Cauchy nel Capitolo 12, mentre i poligoni non regolari appariranno in relazione con il teorema di Pick nel Capitolo 11 e di nuovo quando discuteremo il teorema della galleria d'arte nel Capitolo 31.

I politopi convessi si possono definire in modo alternativo come gli insiemi delle soluzioni limitate di sistemi finiti di disequazioni lineari. Pertanto ogni politopo convesso $P \subseteq \mathbb{R}^d$ ha una rappresentazione della forma

$$P \;=\; \{\boldsymbol{x} \in \mathbb{R}^d : A\boldsymbol{x} \leq \boldsymbol{b}\}$$

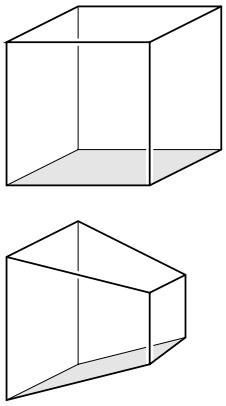
dove $A \in \mathbb{R}^{m \times d}$ è una matrice e $\boldsymbol{b} \in \mathbb{R}^m$ un vettore. In altre parole, P è l'insieme delle soluzioni di un sistema di m disequazioni lineari

$\boldsymbol{a}_i^T \boldsymbol{x} \leq b_i$, dove $\boldsymbol{a}_i^T$ è la riga i-esima di A. Simmetricamente, ogni insieme limitato di soluzioni lineari siffatto è un politopo convesso e pertanto si può rappresentare come l'inviluppo convesso di un insieme finito di punti.

Per poligoni e poliedri abbiamo i concetti familiari di *vertici*, *spigoli* e *bifacce*. Per politopi convessi di dimensioni superiori possiamo definire le facce come segue: una *faccia* di P è un sottoinsieme $F \subseteq P$ della forma $P \cap \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}^T \boldsymbol{x} = b\}$, dove $\boldsymbol{a}^T \boldsymbol{x} \leq b$ è una disequazione lineare valida per tutti i punti $\boldsymbol{x} \in P$. Tutte le facce di un politopo sono esse stesse politopi. L'insieme V dei vertici (facce 0-dimensionali) di un politopo convesso è anche l'insieme di minima inclusione tale che $\text{conv}(V) = P$. Supponendo che $P \subseteq \mathbb{R}^d$ sia un politopo convesso d-dimensionale, le *faccette* (le facce $(d-1)$-dimensionali) determinano un insieme minimale di iperpiani e pertanto di semispazi che contengono P e la cui intersezione è P. In particolare, questo implica il seguente fatto che utilizzeremo in seguito: sia F una faccia di P, si definisca con H_F l'iperpiano che essa determina e con H_F^+ e H_F^- i due semispazi chiusi limitati da H_F. Allora uno di questi due semispazi contiene P (e l'altro non lo contiene).

Il *grafo* $G(P)$ del politopo convesso P è dato dall'insieme V dei vertici e dall'insieme degli spigoli di facce 1-dimensionali E. Se P ha dimensione 3, allora questo grafo è planare, e dà luogo alla famosa "formula del poliedro di Eulero" (si veda il Capitolo 11).

Due politopi $P, P' \subseteq \mathbb{R}^d$ sono *congruenti* se vi è una trasformazione affine che conservi la lunghezza e che trasformi P in P'. Tale trasformazione può invertire l'orientamento dello spazio, come fa la riflessione di P in un iperpiano, che trasforma P in un' *immagine speculare* di P. Essi sono *combinatorialmente equivalenti* se vi è una biiezione fra le facce di P e le facce di P' che conservi la dimensione e le inclusioni tra le facce. Questa nozione di equivalenza combinatoria è molto più debole della congruenza: per esempio, la nostra figura mostra un cubo unitario ed un cubo "deformato" combinatorialmente equivalenti (e pertanto potremmo chiamarli entrambi "cubi"), ma essi non sono certamente congruenti.



Politopi combinatorialmente equivalenti

Un politopo (o un sottoinsieme più generale di $\mathbb{R}^d$) è detto *a simmetria centrale* se esiste un punto $\boldsymbol{x}_0 \in \mathbb{R}^d$ tale che

$$\boldsymbol{x}_0 + \boldsymbol{x} \in P \iff \boldsymbol{x}_0 - \boldsymbol{x} \in P.$$

In questo caso chiamiamo $\boldsymbol{x}_0$ il *centro* di P.

Bibliografia

[1] V. G. BOLTIANSKII: *Hilbert's Third Problem,* V. H. Winston & Sons (Halsted Press, John Wiley & Sons), Washington DC 1978.

[2] M. DEHN: *Ueber raumgleiche Polyeder,* Nachrichten von der Königl. Gesellschaft der Wissenschaften, Mathematisch-physikalische Klasse (1900), 345-354.

[3] M. DEHN: *Ueber den Rauminhalt,* Mathematische Annalen **55** (1902), 465-478.

[4] C. F. GAUSS: *"Congruenz und Symmetrie": Briefwechsel mit Gerling,* pp. 240-249 in: Werke, Band VIII, Königl. Gesellschaft der Wissenschaften zu Göttingen; B. G. Teubner, Leipzig 1900.

[5] D. HILBERT: *Mathematical Problems,* Lecture delivered at the International Congress of Mathematicians at Paris in 1900, Bulletin Amer. Math. Soc. **8** (1902), 437-479.

[6] G. M. ZIEGLER: *Lectures on Polytopes,* Graduate Texts in Mathematics **152**, Springer-Verlag, New York 1995/1998.

Rette nel piano e scomposizioni di grafi

Capitolo 9

Il problema forse più noto sulla configurazione di rette fu sollevato da Sylvester nel 1893 in una raccolta di problemi matematici.

QUESTIONS FOR SOLUTION.

11851. (Professor SYLVESTER.)—Prove that it is not possible to arrange any finite number of real points so that a right line through every two of them shall pass through a third, unless they all lie in the same right line.

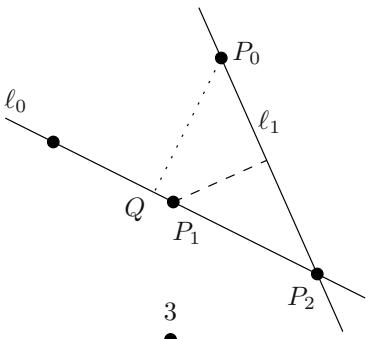
È incerto se Sylvester stesso avesse una dimostrazione, ma una dimostrazione corretta fu data da Tibor Gallai [Grünwald] circa 40 anni più tardi. Pertanto il teorema seguente è comunemente attribuito a Sylvester e Gallai. In seguito alla dimostrazione di Gallai ne apparvero diverse altre, ma il seguente argomento dovuto a L. M. Kelly può essere considerato "semplicemente il migliore".

J. J. Sylvester

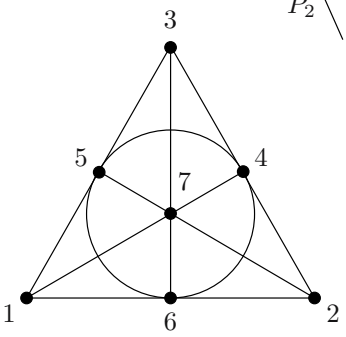
Teorema 1. *In una qualsiasi configurazione di n punti nel piano, non tutti allineati, esiste una retta che contiene esattamente due di questi punti.*

■ **Dimostrazione.** Sia $\mathcal{P}$ l'insieme dei punti dati e si consideri l'insieme $\mathcal{L}$ di tutte le rette che passano per almeno due punti di $\mathcal{P}$. Tra tutte le coppie (P, ℓ) tali che P non si trova su ℓ, si scelga una coppia (P_0, ℓ_0) tale che P_0 abbia distanza minima da ℓ_0 e tale che Q sia il punto su ℓ_0 più vicino a P_0 (ovvero sulla retta passante per P_0 perpendicolare a ℓ_0).

Proposizione. *Questa retta ℓ_0 è quella che cerchiamo!*



Se non lo fosse, allora ℓ_0 conterrebbe almeno tre dei punti di $\mathcal{P}$ e pertanto due di essi, diciamo P_1 e P_2, giacerebbero dalla stessa parte di Q. Supponiamo che P_1 giaccia tra Q e P_2, dove P_1 potrebbe coincidere con Q. La figura a destra mostra la configurazione. Ne segue che la distanza di P_1 dalla retta ℓ_1 determinata da P_0 e da P_2 è minore della distanza da P_0 a ℓ_0, e ciò contraddice la nostra scelta di ℓ_0 e P_0. □



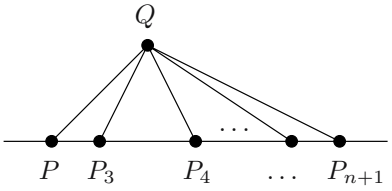
Nella dimostrazione abbiamo usato assiomi metrici (distanza più breve) ed assiomi di ordine (P_1 giace tra Q e P_2) nel piano reale. Abbiamo veramente bisogno di queste proprietà al di là dei consueti assiomi di incidenza di punti e rette? In effetti, c'è bisogno di qualche condizione addizionale, come mostra il celebre piano di Fano rappresentato a margine. Qui $\mathcal{P} = \{1, 2, \ldots, 7\}$ e $\mathcal{L}$ consiste nelle 7 rette individuate da tre punti come indicato nella figura, compresa la "retta" $\{4, 5, 6\}$. Qualunque coppia di punti determina un'unica retta, così gli assiomi di incidenza sono soddisfatti, ma non c'è alcuna retta individuata da solo 2 punti. Il teorema di

Sylvester-Gallai mostra pertanto che la configurazione di Fano non può essere inserita nel piano reale in modo che le sette triplette colineari giacciano su rette reali: ci deve sempre essere una linea spezzata.

Tuttavia, fu mostrato da Coxeter che gli assiomi di ordine sono sufficienti per dimostrare il teorema di Sylvester-Gallai. Pertanto si può produrre una dimostrazione che non faccia uso di alcuna proprietà metrica — si veda anche la dimostrazione che daremo nel Capitolo 11, la quale utilizza la formula di Eulero.

Il teorema di Sylvester-Gallai implica direttamente un altro celebre risultato sui punti e sulle rette nel piano, ad opera di Erdős e Nicolaas G. de Bruijn. Ma in questo caso il risultato vale più generalmente per sistemi arbitrari di punti e rette, come fu osservato già da Erdős e de Bruijn. Discuteremo il risultato più generale in seguito.

Teorema 2. *Sia $\mathcal{P}$ un insieme di $n \geq 3$ punti nel piano, non tutti allineati. Allora l'insieme $\mathcal{L}$ delle rette che passano per almeno due punti contiene almeno n rette.*



■ **Dimostrazione.** Per $n = 3$ non vi è nulla da dimostrare. Ora procediamo per induzione su n. Sia $|\mathcal{P}| = n + 1$. Secondo il teorema precedente esiste una retta $\ell_0 \in \mathcal{L}$ contenente esattamente due punti P e Q di $\mathcal{P}$. Si considerino l'insieme $\mathcal{P}' = \mathcal{P}\backslash\{Q\}$ e l'insieme $\mathcal{L}'$ delle rette determinate da $\mathcal{P}'$. Se i punti di $\mathcal{P}'$ non giacciono tutti su una stessa retta, allora per induzione $|\mathcal{L}'| \geq n$ e pertanto $|\mathcal{L}| \geq n + 1$ a causa della retta addizionale ℓ_0 in $\mathcal{L}$. Se, d'altro canto, i punti di $\mathcal{P}'$ sono tutti allineati, allora abbiamo un "fascio" che si compone esattamente di $n + 1$ rette. □

Ora, come promesso, ecco il risultato generale, che si applica alle molto più generali "geometrie d'incidenza".

Teorema 3. *Sia X un insieme di $n \geq 3$ elementi e siano $A_1, \ldots, A_m$ dei sottoinsiemi propri di X, tali che ogni coppia di elementi di X sia contenuta esattamente in un insieme A_i. Allora si deve avere $m \geq n$.*

■ **Dimostrazione.** La dimostrazione seguente, attribuita ora a Motzkin ora a Conway, è lunga poco più di una riga e veramente ispirata. Per $x \in X$ sia r_x il numero di insiemi A_i contenenti x. (Si noti che $2 \leq r_x < m$ grazie alle ipotesi). Ora se $x \notin A_i$, allora $r_x \geq |A_i|$ poiché gli insiemi $|A_i|$ contententi x ed un elemento di A_i debbono essere distinti. Supponiamo che $m < n$, allora $m|A_i| < n\, r_x$ e pertanto $m(n - |A_i|) > n(m - r_x)$ per $x \notin A_i$ e troviamo

$$1 = \sum_{x\in X} \tfrac{1}{n} = \sum_{x\in X} \sum_{A_i : x\notin A_i} \tfrac{1}{n(m-r_x)} > \sum_{A_i} \sum_{x : x\notin A_i} \tfrac{1}{m(n-|A_i|)} = \sum_{A_i} \tfrac{1}{m} = 1,$$

il che è assurdo. □

Vi è un'altra dimostrazione molto breve di questo teorema che fa uso dell'algebra lineare. Sia B la *matrice d'incidenza* di $(X; A_1, \ldots, A_m)$, ovvero le righe in B sono indicizzate dagli elementi di X e le colonne da

$A_1, \ldots, A_m$, dove

$$B_{xA} := \begin{cases} 1 & \text{se} \quad x \in A \\ 0 & \text{se} \quad x \notin A. \end{cases}$$

Si consideri il prodotto BB^T. Per $x \neq x'$ abbiamo $(BB^T)_{xx'} = 1$, poiché x e x' sono contenuti precisamente in un insieme A_i, si ha

$$BB^T = \begin{pmatrix} r_{x_1}-1 & 0 & \ldots & 0 \\ 0 & r_{x_2}-1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & r_{x_n}-1 \end{pmatrix} + \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & & \vdots \\ \vdots & & \ddots & 1 \\ 1 & \ldots & 1 & 1 \end{pmatrix}$$

dove r_x è definito come sopra. Poiché la prima matrice è definita positiva (ha solo autovalori positivi) e la seconda matrice è semi-definita positiva (ha come autovalori n e 0), deduciamo che BB^T è definita positiva e pertanto, in particolare, è invertibile, il che implica che il suo rango è n. Ne segue che il rango della matrice B $(n \times m)$ è almeno n e concludiamo che in effetti $n \leq m$, poiché il rango non può superare il numero delle colonne.

Andiamo un po' oltre e consideriamo la teoria dei grafi. (Facciamo riferimento al riassunto di concetti basilari sui grafi che si trova nell'appendice a questo capitolo). Una breve riflessione mostra che l'affermazione seguente è in verità identica al Teorema 3:

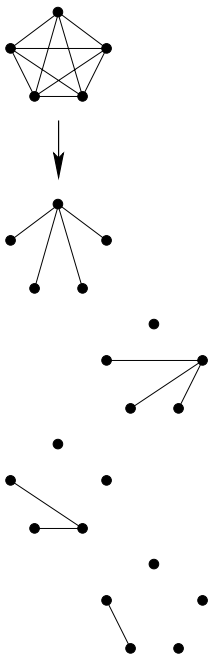
> *Se scomponiamo un grafo completo K_n in m clique diverse da K_n, tali che ogni arco sia in un'unica clique, allora $m \geq n$.*

In effetti, facendo corrispondere X con l'insieme dei vertici di K_n e gli insiemi A_i con gli insiemi dei vertici delle clique, le due formulazioni sono identiche.

Il nostro prossimo obiettivo è di scomporre K_n in grafi completi bipartiti tali che ancora una volta ogni arco sia esattamente in uno di questi grafi. Vi è un modo semplice di svolgere questa operazione. Si numerino i vertici $\{1, 2, \ldots, n\}$. Per prima cosa, si prenda il grafo bipartito completo che congiunge 1 con tutti gli altri vertici. Otteniamo così il grafo $K_{1,n-1}$, detto *stella*. In seguito si congiunga 2 con $3, \ldots, n$, ottenendo come risultato una stella $K_{1,n-2}$. Continuando così, scomponiamo K_n in stelle $K_{1,n-1}, K_{1,n-2}, \ldots, K_{1,1}$. Questa scomposizione usa $n-1$ grafi bipartiti completi. Possiamo fare di meglio, ovvero utilizzare meno grafi? No, come afferma il seguente risultato di Ron Graham e Henry O. Pollak:

Teorema 4. *Se K_n è scomposto in sottografi bipartiti completi $H_1, \ldots, H_m$, allora $m \geq n-1$.*

È interessante osservare che, contrariamente al teorema di Erdős-de Bruijn, non è nota alcuna dimostrazione combinatoria di questo teorema! Tutte usano l'algebra lineare, in un modo o nell'altro. Fra le varie idee più o meno equivalenti consideriamo qui la dimostrazione di Tverberg, che è probabilmente la più chiara.



Una scomposizione di K_5 in 4 sottografi bipartiti completi

■ **Dimostrazione.** Sia $\{1, \ldots, n\}$ l'insieme dei vertici di K_n, e siano L_j, R_j gli insiemi che definiscono i vertici del grafo bipartito completo H_j, $j = 1, \ldots, m$. Associamo ad ogni vertice i una variabile x_i. Poiché $H_1, \ldots, H_m$ scompongono K_n, troviamo

$$\sum_{i<j} x_i x_j \;=\; \sum_{k=1}^{m} \Big(\sum_{a \in L_k} x_a \cdot \sum_{b \in R_k} x_b \Big). \tag{1}$$

Supponiamo ora che il teorema sia falso, ovvero che $m < n - 1$. Allora il sistema di equazioni lineari

$$\begin{aligned} x_1 + \ldots + x_n &= 0, \\ \sum_{a \in L_k} x_a &= 0 \qquad (k = 1, \ldots, m) \end{aligned}$$

ha meno equazioni che incognite, pertanto esiste una soluzione non banale $c_1, \ldots, c_n$. Da (1) deduciamo

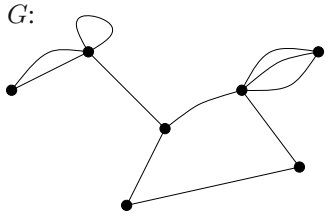
$$\sum_{i<j} c_i c_j \;=\; 0.$$

Ma ciò conduce alla contraddizione

$$0 \;=\; (c_1 + \ldots + c_n)^2 \;=\; \sum_{i=1}^{n} c_i^2 + 2 \sum_{i<j} c_i c_j \;=\; \sum_{i=1}^{n} c_i^2 \;>\; 0,$$

e la dimostrazione è completa. □

Appendice: concetti elementari sui grafi



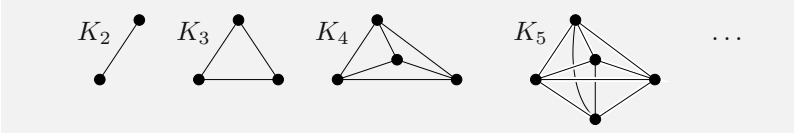
Un grafo G con 7 vertici e 11 archi. Esso ha un ciclo, un arco doppio ed uno triplo

I grafi sono fra le strutture matematiche più basilari. Per questo, essi hanno differenti versioni, molte rappresentazioni ed incarnazioni. In astratto, un *grafo* è una coppia $G = (V, E)$, dove V è l'insieme di *vertici*, E è l'insieme di *archi* ed ogni arco $e \in E$ "collega" due vertici $v, w \in V$. Consideriamo soltanto grafi finiti, dove V e E sono finiti.

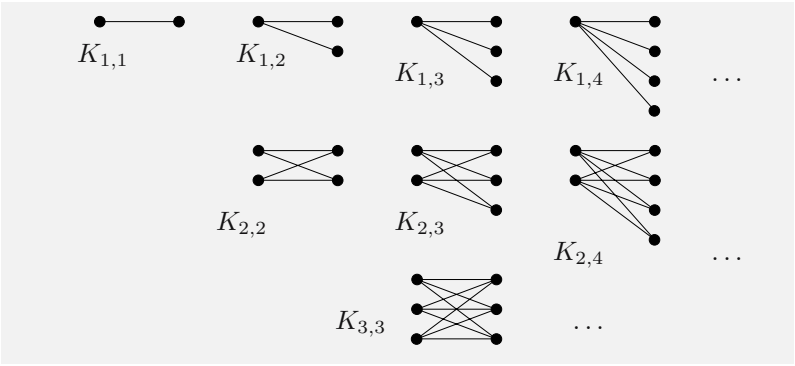
Generalmente abbiamo a che fare con *grafi semplici*: in questo caso non ammettiamo *cicli*, i. e. archi per i quali entrambe le terminazioni coincidono, e neppure *archi multipli* che hanno lo stesso insieme di vertici come terminazioni.

I vertici di un grafo sono detti *adiacenti* o *vicini* se sono gli estremi di un arco. Un vertice ed un arco sono detti *incidenti* se l'arco ha quel vertice come estremo.

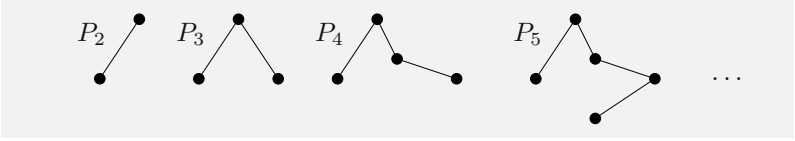
Ecco una piccola immagine che rappresenta grafi (semplici) importanti:



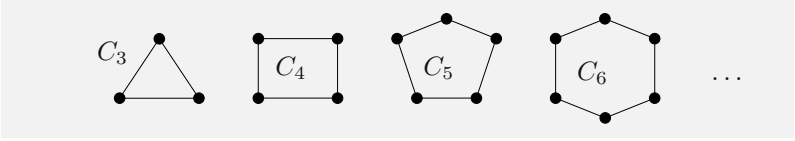
I *grafi completi* K_n su n vertici e $\binom{n}{2}$ archi



I *grafi bipartiti completi* $K_{m,n}$ con $m+n$ vertici e mn archi



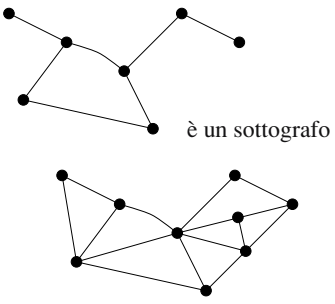
I *cammini* P_n con n vertici



I *cicli* C_n con n vertici

Due grafi $G = (V, E)$ e $G' = (V', E')$ sono considerati *isomorfi* se esistono delle biiezioni $V \to V'$ e $E \to E'$ che conservino le incidenze tra gli archi ed i loro estremi. (È un problema tuttora irrisolto se esista un test efficace per decidere se due grafi dati siano isomorfi). La nozione di isomorfismo ci permette di parlare *del* grafo completo K_5 su 5 vertici, etc.

$G' = (V', E')$ è un *sottografo* di $G = (V, E)$ se $V' \subseteq V, E' \subseteq E$, ed ogni arco $e \in E'$ ha gli stessi estremi in G' ed in G. G' è un *sottografo indotto* se, in aggiunta, *tutti* gli archi di G che collegano i vertici di G' sono anch'essi archi di G'.



Molte nozioni sui grafi sono alquanto intuitive: per esempio, un grafo G è *connesso* se ogni coppia di vertici distinti è connessa da un cammino in G, o equivalentemente se G non può essere diviso in due sottografi non vuoti i cui insiemi di vertici siano disgiunti.

Terminiamo questa breve rassegna dei concetti di base sui grafi con qualche altra definizione: una *clique* in G è un sottografo completo. Un *insieme indipendente* in G è un sottografo indotto senza archi, ovvero un sottoinsieme dell'insieme di vertici tale che nessuna coppia di vertici sia connessa da un arco di G. Un grafo è una *foresta* se non contiene cicli. Un *albero*

è una foresta connessa. Infine, un grafo $G = (V, E)$ è *bipartito* se è isomorfo ad un sottografo di un grafo completo bipartito ovvero se l'insieme dei suoi vertici si può scrivere come l'unione $V = V_1 \cup V_2$ di due insiemi indipendenti.

Bibliografia

[1] N. G. DE BRUIJN & P. ERDŐS: *On a combinatorial problem,* Proc. Kon. Ned. Akad. Wetensch. **51** (1948), 1277-1279.

[2] H. S. M. COXETER: *A problem of collinear points,* Amer. Math. Monthly **55** (1948), 26-28 (contains Kelly's proof).

[3] P. ERDŐS: *Problem 4065 — Three point collinearity,* Amer. Math. Monthly **51** (1944), 169-171 (contains Gallai's proof).

[4] R. L. GRAHAM & H. O. POLLAK: *On the addressing problem for loop switching,* Bell System Tech. J. **50** (1971), 2495-2519.

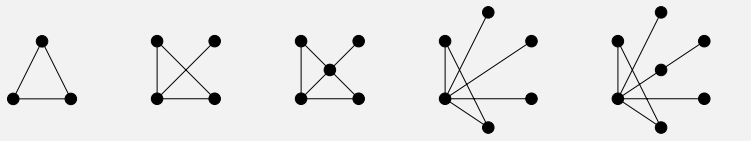
[5] J. J. SYLVESTER: *Mathematical Question 11851,* The Educational Times **46** (1893), 156.

[6] H. TVERBERG: *On the decomposition of K_n into complete bipartite graphs,* J. Graph Theory **6** (1982), 493-494.

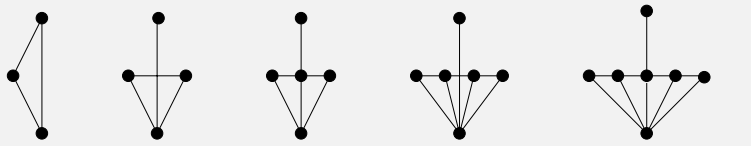
Il problema delle pendenze

Capitolo 10

Provate voi stessi — prima di continuare la lettura di questo capitolo — a costruire configurazioni di punti nel piano che determinino "relativamente poche" pendenze. Supponiamo, naturalmente, che gli $n \geq 3$ punti non giacciano tutti sulla stessa retta. Si ricordi dal Capitolo 9 il teorema di Erdős e de Bruijn: gli n punti determineranno almeno n rette distinte. Ma ovviamente molte di queste rette potrebbero essere parallele e pertanto avere la stessa pendenza.



ovvero

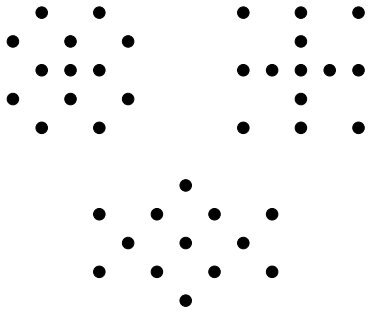


Un rapido esperimento per n piccoli vi porterà probabilmente ad una delle sequenze qui riportate

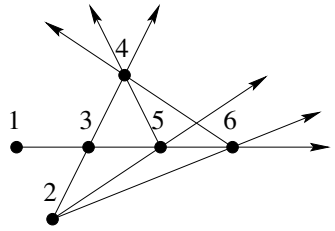
Dopo alcuni tentativi di trovare configurazioni con meno pendenze potreste congetturare — come fece Scott nel 1970 — il seguente teorema.

Teorema. *Se $n \geq 3$ punti nel piano non giacciono su una stessa retta, allora essi determinano almeno $n - 1$ pendenze diverse, dove l'uguaglianza è possibile solo se n è dispari e $n \geq 5$.*

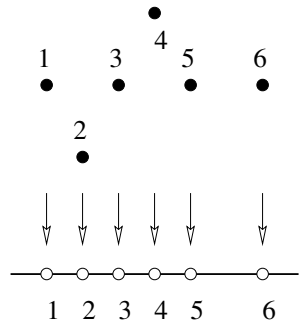
Gli esempi riportati sopra — i disegni rappresentano le prime configurazioni di due successioni infinite di esempi — mostrano che il teorema come enunciato è *il migliore possibile*: per ogni $n \geq 5$ dispari vi è una configurazione con n punti che determina esattamente $n - 1$ pendenze diverse e per



Tre begli esempi sporadici dal catalogo di Jamison-Hill



Questa configurazione di $n = 6$ punti determina $t = 6$ pendenze diverse



Qui una direzione di partenza verticale dà $\pi_0 = 123456$

un qualunque altro $n \geq 3$ abbiamo una configurazione con esattamente n pendenze.

Tuttavia, le configurazioni che abbiamo disegnato sopra sono lungi dall'essere le uniche. Per esempio, Jamison e Hill descrissero quattro famiglie infinite di configurazioni, ognuna delle quali consisteva in configurazioni con un numero dispari n di punti che determinavano soltanto $n-1$ pendenze ("configurazioni di critica pendenza"). Inoltre, essi enumerarono 102 esempi sporadici che non sembrano rientrare in una famiglia infinita, la maggior parte dei quali trovati grazie a ricerche intensive fatte con l'uso del calcolatore.

Il buon senso potrebbe portarci ad affermare che i problemi estremali tendono ad essere molto difficili da risolvere proprio quando le configurazioni estreme sono così diverse ed irregolari. In effetti, si può dire molto sulla struttura delle configurazioni di pendenza critica (si veda [2]), ma una classificazione sembra completamente fuori portata. Tuttavia, il teorema precedente ha una dimostrazione semplice, che è costituita da due ingredienti principali: la riduzione ad un modello combinatorio semplice dovuta a Eli Goodman and Ricky Pollack ed una splendida argomentazione in questo modello con cui Peter Ungar completò la dimostrazione nel 1982.

■ **Dimostrazione. (1)** Innanzitutto notiamo che basta mostrare che ogni insieme pari di $n = 2m$ punti nel piano $(m \geq 2)$ determina almeno n pendenze. Questo poiché il caso $n = 3$ è banale e per ogni insieme di $n = 2m+1 \geq 5$ punti (non tutti allineati) possiamo trovare un sottoinsieme di $n-1 = 2m$ punti, non tutti allineati, che determina già $n-1$ pendenze.

Pertanto nel seguito consideriamo una configurazione di $n = 2m$ punti nel piano che determina $t \geq 2$ pendenze diverse.

(2) Il modello combinatorio si ottiene costruendo una sequenza periodica di permutazioni. A questo scopo cominciamo con una qualche direzione nel piano delle permutazioni che non sia una della pendenze della configurazione e numeriamo i punti $1, \ldots, n$ nell'ordine nel quale appaiono nella proiezione monodimensionale in questa direzione. Pertanto la permutazione $\pi_0 = 123...n$ rappresenta l'ordine dei punti per la nostra direzione di partenza.

Successivamente facciamo compiere alla direzione un movimento in senso anti-orario e osserviamo come cambiano la proiezione e la sua permutazione. I cambiamenti nell'ordine dei punti proiettati appaiono esattamente quando la direzione passa su di una delle pendenze della configurazione.

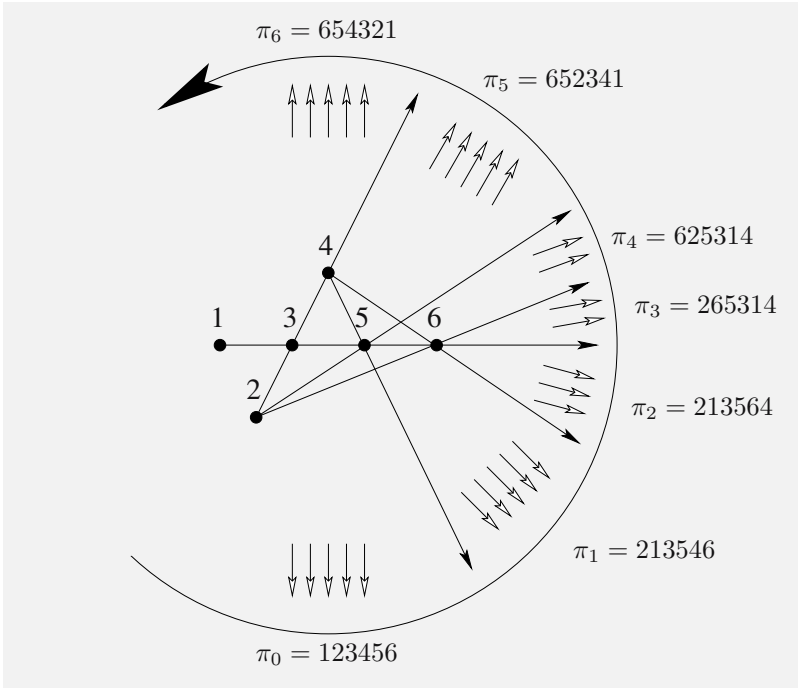
Tuttavia i cambiamenti sono tutt'altro che casuali o arbitrari: compiendo una rotazione della direzione di $180°$ otteniamo una sequenza di permutazioni

$$\pi_0 \rightarrow \pi_1 \rightarrow \pi_2 \rightarrow \ldots \rightarrow \pi_{t-1} \rightarrow \pi_t$$

che ha le seguenti proprietà speciali:

- La sequenza comincia con $\pi_0 = 123...n$ e termina con $\pi_t = n...321$.
- La lunghezza t della sequenza è il numero di pendenze della configurazione di punti.

- Lungo la sequenza, ogni coppia $i < j$ è scambiata esattamente una volta. Questo significa che nel percorso da $\pi_0 = 123...n$ a $\pi_t = n...321$, vengono invertite solo sottostringhe *crescenti*.
- Ogni movimento consiste nell'inversione di una o più sottostringhe crescenti disgiunte (che corrispondono ad una o più rette parallele alla direzione data).



Come ottenere la sequenza delle permutazioni per il nostro piccolo esempio

Continuando il movimento circolare intorno alla configurazione, si può vedere la sequenza come parte di una sequenza infinita e bidirezionale di permutazioni

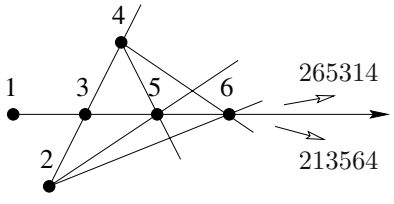
$$\ldots \to \pi_{-1} \to \pi_0 \to \ldots \to \pi_t \to \pi_{t+1} \to \ldots \to \pi_{2t} \to \ldots$$

dove π_{i+t} è l'inverso di π_i per ogni i, e pertanto $\pi_{i+2t} = \pi_i$ per ogni $i \in \mathbb{Z}$.

Mostreremo che *ogni* sequenza con le proprietà sopracitate (e $t \geq 2$) deve avere lunghezza $t \geq n$.

(3) La chiave della dimostrazione è dividere ogni permutazione in una "metà sinistra" ed una "metà destra" di uguale dimensione $m = \frac{n}{2}$ e di contare le lettere che incrociano la *barriera* immaginaria tra la metà destra e quella sinistra.

Chiamiamo $\pi_i \to \pi_{i+1}$ un movimento *attraversante* se una delle sottocatene che esso inverte coinvolge lettere da entrambe le parti della barriera. Il movimento attraversante ha *ordine* d se esso sposta $2d$ lettere attraverso la barriera, ovvero, se la catena che incrocia ha esattamente d lettere da una parte ed almeno d lettere dall'altra. Pertanto nel nostro esempio



Un movimento attraversante

$$\pi_2 \;=\; 2\underline{13}{:}\underline{56}4 \longrightarrow 2\overline{65{:}31}4 \;=\; \pi_3$$

è un movimento attraversante di ordine $d = 2$ (sposta $1, 3, 5, 6$ attraverso la barriera, che indichiamo con ":"),

$$65\underline{2}{:}\underline{34}1 \longrightarrow 65\overline{4{:}32}1$$

è un movimento attraversante di ordine $d = 1$, mentre per esempio

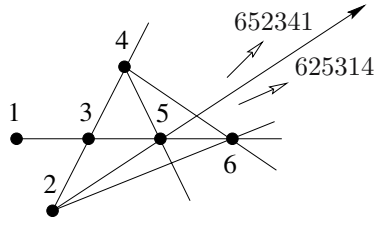
$$6\underline{25}{:}3\underline{14} \longrightarrow 6\overline{52}{:}3\overline{41}$$

non è un movimento attraversante.

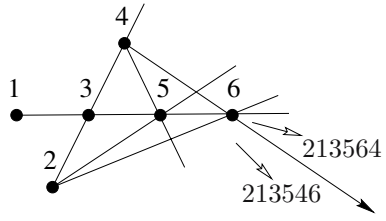
Nel corso della sequenza $\pi_0 \to \pi_1 \to \ldots \to \pi_t$, ognuna delle lettere $1, 2, \ldots, n$ deve incrociare la barriera almeno una volta. Ciò implica che, se gli ordini dei c movimenti attraversanti sono $d_1, d_2, \ldots, d_c$, allora abbiamo

$$\sum_{i=1}^{c} 2d_i \;=\; \#\,\{\text{lettere che attraversano la barriera}\} \;\geq\; n.$$

Pertanto abbiamo almeno due movimenti di attraversamento, poiché un movimento di attraversamento con $2d_i = n$ avviene soltanto se tutti i punti sono allineati, i. e. per $t = 1$. Geometricamente, un movimento attraversante corrisponde alla direzione di una retta nella configurazione che ha meno di m punti da ogni parte.



Un movimento toccante



Un movimento ordinario

(4) Un *movimento toccante* è un movimento che inverte una catena adiacente alla barriera centrale, ma non la incrocia. Per esempio

$$\pi_4 \;=\; 6\underline{25}{:}3\underline{14} \longrightarrow 6\overline{52}{:}3\overline{41} \;=\; \pi_5$$

è un movimento toccante. Geometricamente, un movimento toccante corrisponde alla pendenza di una retta della configurazione che ha esattamente m punti da una parte e pertanto al massimo $m - 2$ punti dall'altro lato.

I movimenti che non sono né toccanti né attraversanti saranno chiamati *movimenti ordinari*. Un esempio di questo è

$$\pi_1 \;=\; 213{:}5\underline{46} \longrightarrow 213{:}5\overline{64} \;=\; \pi_2.$$

Quindi ogni movimento è o attraversante o toccante o ordinario e useremo le lettere C, T, O per indicare i tipi di movimento (NdT: C sta per *crossing*, il termine inglese per "attraversante"). $C(d)$ indicherà un movimento attraversante di ordine d. Pertanto per il nostro piccolo esempio abbiamo

$$\pi_0 \xrightarrow{T} \pi_1 \xrightarrow{O} \pi_2 \xrightarrow{C(2)} \pi_3 \xrightarrow{O} \pi_4 \xrightarrow{T} \pi_5 \xrightarrow{C(1)} \pi_6;$$

ancora più brevemente possiamo registrare questa sequenza come $T, O, C(2), O, T, C(1)$.

(5) Per completare la dimostrazione, abbiamo bisogno dei due proprietà seguenti:

> *Tra due movimenti attraversanti vi è almeno un movimento toccante.*
>
> *Tra qualunque movimento attraversante di ordine d ed il successivo movimento toccante vi sono almeno $d-1$ movimenti ordinari.*

In effetti, dopo un movimento attraversante di ordine d la barriera è contenuta in una sottocatena decrescente simmetrica di lunghezza $2d$, con d lettere da ogni parte della barriera. Per il movimento attraversante successivo la barriera centrale deve essere portata in una sottostringa crescente di lunghezza almeno 2. Ma solo i movimenti toccanti influenzano il fatto che la barriera sia una sottostringa crescente. Questo dà la prima proprietà. Per il secondo, si noti che con ogni movimento ordinario (che incroci qualche sottocatena *crescente*) la $2d$-catena decrescente può essere accorciata di una lettera soltanto da ogni parte. Inoltre, fintanto che la catena decrescente ha almeno 4 lettere, un movimento attraversante è impossibile. Questo dà la seconda proprietà.

Se costruiamo la sequenza di permutazioni che inizia con la stessa proiezione iniziale ma che opera una rotazione in senso orario, allora otteniamo la sequenza invertita di permutazioni. Pertanto la sequenza che abbiamo registrato deve anche soddisfare il contrario della nostra seconda affermazione, ovvero

> *Tra un movimento attraversante ed il successivo movimento attraversante, di ordine d, vi sono almeno $d-1$ movimenti ordinari.*

(6) Il *modello* T-O-C dell'infinita sequenza di permutazioni, come derivato in **(2)**, si ottiene ripetendo continuamente il motivo T-O-C di lunghezza t della sequenza $\pi_0 \longrightarrow \ldots \longrightarrow \pi_t$. Pertanto con le affermazioni in **(5)** vediamo che la sequenza infinita di movimenti, ognuno attraversante e di ordine d, si inserisce in un motivo T-O-C del tipo

$$T, \underbrace{O, O, \ldots, O}_{\geq d-1}, C(d), \underbrace{O, O, \ldots, O}_{\geq d-1}, \qquad (*)$$

di lunghezza $1 + (d-1) + 1 + (d-1) = 2d$.

Nella sequenza infinita, possiamo considerare un segmento finito di lunghezza t che incomincia con un movimento attraversante. Tale segmento consiste in sottocatene del tipo $(*)$, più eventualmente altre T inserite. Ciò implica che la sua lunghezza t soddisfa

$$t \;\geq\; \sum_{i=1}^{c} 2d_i \;\geq\; n,$$

il che completa la dimostrazione. □

Bibliografia

[1] J. E. GOODMAN & R. POLLACK: *A combinatorial perspective on some problems in geometry,* Congressus Numerantium **32** (1981), 383-394.

[2] R. E. JAMISON & D. HILL: *A catalogue of slope-critical configurations,* Congressus Numerantium **40** (1983), 101-125.

[3] P. R. SCOTT: *On the sets of directions determined by n points,* Amer. Math. Monthly **77** (1970), 502-505.

[4] P. UNGAR: *$2N$ noncollinear points determine at least $2N$ directions,* J. Combinatorial Theory, Ser. A **33** (1982), 343-347.

Tre applicazioni della formula di Eulero

Capitolo 11

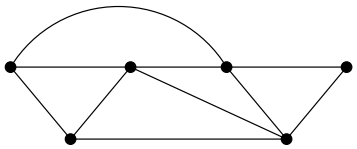
Un grafo è detto *planare* se può essere disegnato nel piano $\mathbb{R}^2$ (o, equivalentemente, sulla sfera di dimensione S^2) senza che si debbano attraversare archi. Parliamo di grafo *piano* se tale disegno è già dato e fissato. Qualunque disegno di questo tipo scompone il piano o la sfera in un numero finito di regioni connesse, compresa la regione esterna (illimitata), indicate col termine *facce*. La formula di Eulero mostra una splendida relazione tra il numero di vertici, archi e facce valida per ogni grafo piano. Eulero menzionò questo risultato per la prima volta in una lettera al suo amico Goldbach nel 1750, ma all'epoca non disponeva di una dimostrazione completa. Tra le molte dimostrazioni della formula di Eulero, ne presentiamo una bella e auto-duale che non fa uso del principio di induzione. Essa si può far risalire all'opera di von Staudt "Geometrie der Lage" del 1847.

Eulero

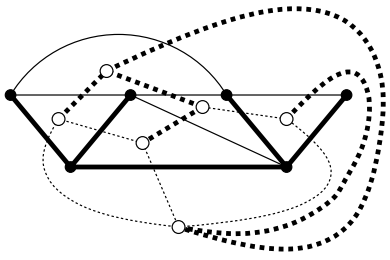
> **Formula di Eulero.** *Se G è un grafo piano connesso con n vertici, e archi e f facce, allora*
>
> $$n - e + f \;=\; 2.$$

■ **Dimostrazione.** Sia $T \subseteq E$ l'insieme degli archi di un albero generatore (NdT: spanning tree nell'edizione inglese) per G, ovvero di un sottografo minimo che connetta tutti i vertici di G. Tale grafo non contiene cicli a causa dell'ipotesi di minimalità.

Ora ci serve il *grafo duale* G^* di G: per costruirlo, si metta un vertice all'interno di ogni faccia di G e si colleghino due vertici così costruiti di G^* con archi che corrispondano ad archi di bordo comuni alle facce corrispondenti. Se ci sono diversi archi di bordo comuni, allora disegnamo diversi archi di collegamento nel grafo duale. (Pertanto G^* può avere archi multipli anche se il grafo originale G è semplice.)

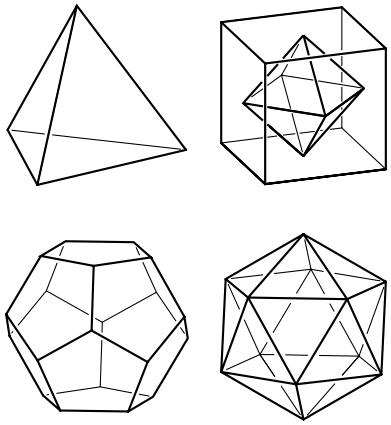

Un grafo piano G: $n = 6$, $e = 10$, $f = 6$

Si consideri la collezione $T^* \subseteq E^*$ di archi nel grafo duale corrispondente agli archi in $E \backslash T$. Gli archi in T^* collegano tutte le facce, poiché T non ha alcun ciclo; ma anche T^* non contiene alcun ciclo, poiché altrimenti esso separerebbe i vertici di G all'interno del ciclo dai vertici all'esterno (e ciò non può essere, poiché T è un sottografo generatore e non vi è intersezione tra gli archi di T e quelli di T^*). Pertanto T^* è un albero generatore per G^*.


Alberi generatori duali in G ed in G^*

Per ogni albero il numero di vertici supera di uno il numero di archi. Per verificarlo, si scelga un vertice come radice e si dirigano tutti gli archi "lontano dalla radice": ciò dà una biiezione tra i vertici diversi dalla radice e gli archi, associando ogni arco con il vertice verso cui punta. Applicato all'albero T questo ragionamento dà $n = e_T + 1$, mentre per l'albe-

ro T^* dà $f = e_{T^*} + 1$. Sommando le due equazioni otteniamo $n + f = (e_T + 1) + (e_{T^*} + 1) = e + 2$. □

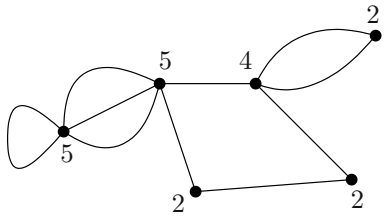


I cinque solidi platonici

La formula di Eulero fornisce dunque una notevole conclusione *numerica* a partire da una situazione *geometrico-topologica*: il numero di vertici, archi e facce di un grafo finito G soddisfa $n - e + f = 2$ ogniqualvolta il grafo sia *o possa essere* disegnato nel piano o su una sfera.

Dalla formula di Eulero si possono trarre molte conseguenze ben note ed ormai classiche. Tra di esse vi sono la classificazione dei poliedri regolari convessi (i solidi platonici), il fatto che K_5 e $K_{3,3}$ non sono planari (si veda sotto) ed il teorema dei cinque colori secondo cui ogni carta piana si può colorare con al massimo cinque colori tali che nessuna regione adiacente abbia lo stesso colore. Ma di questo risultato abbiamo una dimostrazione decisamente migliore, che non necessita neppure della formula di Eulero — si veda al Capitolo 30.

Questo capitolo raccoglie tre altre splendide dimostrazioni che hanno come nucleo la formula di Eulero. Le prime due — una dimostrazione del teorema di Sylvester-Gallai e di un teorema sulle configurazioni bicolori di punti — fanno uso della formula di Eulero combinandola opportunamente con altre relazioni aritmetiche tra parametri basilari dei grafi. Analizziamo dapprima questi parametri.



Qui il grado è scritto di fianco ad ogni vertice. Contando i vertici di grado assegnato abbiamo $n_2 = 3$, $n_3 = 0$, $n_4 = 1$, $n_5 = 2$

Il *grado* di un vertice è il numero di archi che terminano nel vertice, dove i cicli contano doppio. Sia n_i il numero di vertici di grado i in G. Contando i vertici secondo i loro gradi, otteniamo

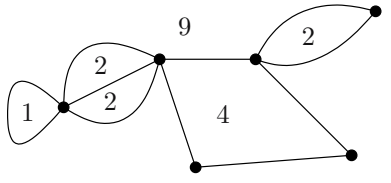
$$n \;=\; n_0 + n_1 + n_2 + n_3 + \dots \tag{1}$$

D'altro canto, ogni arco ha due estremi, pertanto contribuisce con 2 alla somma di tutti i gradi. Di conseguenza otteniamo

$$2e \;=\; n_1 + 2n_2 + 3n_3 + 4n_4 + \dots \tag{2}$$

Questa identità si può interpretare contando in due modi i termini degli archi, ovvero le incidenze arco-vertice. Il *grado medio* $\overline{d}$ dei vertici è pertanto

$$\overline{d} \;=\; \frac{2e}{n}.$$



Il numero di lati è scritto in ogni regione. Contando le facce con un dato numero di lati si ottiene $f_1 = 1$, $f_2 = 3$, $f_4 = 1$, $f_9 = 1$, e $f_i = 0$ altrimenti

In seguito conteremo le facce di un grafo piano secondo il loro numero di lati: una *k-faccia* è una faccia limitata da k archi (dove un arco che limita la stessa regione da entrambi i lati deve essere contato due volte!). Sia f_k il numero di k-facce. Contando tutte le facce troviamo

$$f \;=\; f_1 + f_2 + f_3 + f_4 + \dots \tag{3}$$

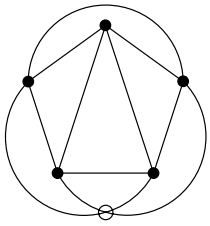
Contando gli archi secondo le facce di cui sono lati, otteniamo

$$2e \;=\; f_1 + 2f_2 + 3f_3 + 4f_4 + \dots \tag{4}$$

Di nuovo, possiamo interpretare questo risultato come un doppio conto delle incidenze archi-facce. Si noti che il numero medio di lati delle facce è dato da

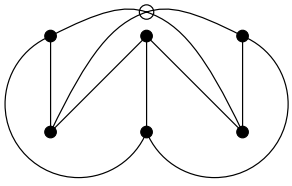
$$\overline{f} \;=\; \frac{2e}{f}.$$

Da questo — e dalla formula di Eulero — deduciamo rapidamente che il grafo completo K_5 ed il grafo completo bipartito $K_{3,3}$ non sono planari. Per un ipotetico disegno piano di K_5 calcoliamo $n = 5$, $e = \binom{5}{2} = 10$, pertanto $f = e + 2 - n = 7$ e $\overline{f} = \frac{2e}{f} = \frac{20}{7} < 3$. Ma se il numero medio di lati fosse inferiore a 3, allora l'inclusione avrebbe una faccia con al massimo due lati, il che è impossibile.



K_5 disegnato con un'intersezione

Analogamente per $K_{3,3}$ otteniamo $n = 6$, $e = 9$ e $f = e + 2 - n = 5$, pertanto $\overline{f} = \frac{2e}{f} = \frac{18}{5} < 4$, il che non può essere dal momento che $K_{3,3}$ è semplice e bipartito, dunque tutti i suoi cicli hanno almeno lunghezza 4.



$K_{3,3}$ disegnato con un'intersezione

Non è una coincidenza, ovviamente, che le equazioni (3) e (4) per gli f_i sembrino così simili alle equazioni (1) e (2) per gli n_i. Esse si trasformano le une nelle altre attraverso la costruzione dei grafi duali $G \to G^*$ spiegata in precedenza.

Dalle precedenti identità otteniamo le seguenti importanti conseguenze "locali" della formula di Eulero.

Proposizione. *Sia G un qualunque grafo piano semplice con $n > 2$ vertici. Allora*

(A) *G ha un vertice di grado al più 5.*

(B) *G ha al più $3n - 6$ archi.*

(C) *Se gli archi di G sono bicolori, allora vi è un vertice di G con al più due cambi di colore nell'ordine ciclico degli archi intorno al vertice.*

■ **Dimostrazione.** Per ognuna delle tre affermazioni, possiamo supporre che G sia connesso.

(A) Ogni faccia ha almeno 3 lati (poiché G è semplice), così (3) e (4) danno

$$\begin{aligned} f &= f_3 + f_4 + f_5 + \dots \\ 2e &= 3f_3 + 4f_4 + 5f_5 + \dots \end{aligned}$$

e pertanto $2e - 3f \geq 0$.

Ora se ogni vertice ha almeno grado 6, allora (1) e (2) implicano

$$\begin{aligned} n &= n_6 + n_7 + n_8 + \dots \\ 2e &= 6n_6 + 7n_7 + 8n_8 + \dots \end{aligned}$$

e pertanto $2e - 6n \geq 0$.

Considerando entrambe le diseguaglianze, otteniamo

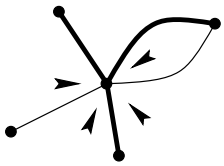
$$6(e - n - f) \;=\; (2e - 6n) + 2(2e - 3f) \;\geq\; 0$$

e pertanto $e \geq n + f$, il che contraddice la formula di Eulero.

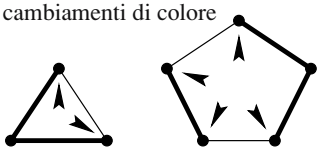
(B) Come nella prima fase della parte (A), otteniamo $2e-3f \geq 0$ e pertanto

$$3n - 6 \;=\; 3e - 3f \;\geq\; e$$

grazie alla formula di Eulero.



Le frecce puntano verso gli angoli senza cambiamenti di colore



(C) Sia c il numero di angoli dove avviene il cambio di colore. Supponiamo che l'affermazione sia falsa: allora abbiamo $c \geq 4n$ angoli con cambi di colore, poiché ad ogni vertice vi è un numero pari di cambi. Ora ogni faccia con $2k$ o $2k+1$ lati ha al massimo $2k$ di tali angoli, così concludiamo che

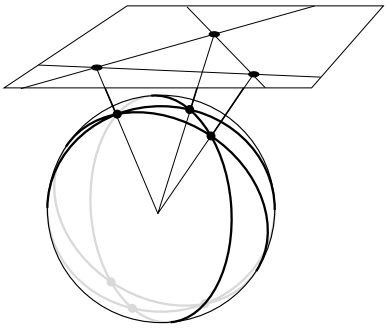
$$\begin{aligned} 4n \;\leq\; c \;&\leq\; 2f_3 + 4f_4 + 4f_5 + 6f_6 + 6f_7 + 8f_8 + \ldots \\ &\leq\; 2f_3 + 4f_4 + 6f_5 + 8f_6 + 10f_7 + \ldots \\ &=\; 2(3f_3 + 4f_4 + 5f_5 + 6f_6 + 7f_7 + \ldots) \\ &\quad\; -4(f_3 + f_4 + f_5 + f_6 + f_7 + \ldots) \\ &=\; 4e - 4f \end{aligned}$$

usando ancora (3) e (4). Abbiamo pertanto $e \geq n + f$, contraddicendo ancora la formula di Eulero. □

1. Il teorema di Sylvester-Gallai rivisitato

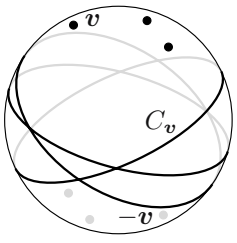
Norman Steenrod fu il primo, a quanto sembra, a notare che la parte (A) della proposizione precedente fornisce una dimostrazione sorprendentemente semplice del teorema di Sylvester-Gallai (si veda il Capitolo 9).

> **Il teorema di Sylvester-Gallai.** *Dato un insieme di $n \geq 3$ punti nel piano, non tutti allineati, vi è sempre una retta che contiene esattamente due dei punti.*



■ **Dimostrazione.** (Sylvester-Gallai via Eulero)
Se inseriamo il piano $\mathbb{R}^2$ in $\mathbb{R}^3$ vicino alla sfera unitaria S^2 come indicato nella nostra figura, allora ogni punto in $\mathbb{R}^2$ corrisponde ad una coppia di punti agli antipodi su S^2, e le rette in $\mathbb{R}^2$ corrispondono a cerchi massimi in S^2. Pertanto il teorema di Sylvester-Gallai si riduce a quanto segue:

> *Dato un qualunque insieme di $n \geq 3$ coppie di punti agli antipodi sulla sfera, non tutti su di un cerchio massimo, vi è sempre un cerchio massimo che contiene esattamente due delle coppie agli antipodi.*



Ora procediamo per dualità, sostituendo ogni coppia di punti agli antipodi con il cerchio massimo corrispondente sulla sfera. Ovvero invece dei punti $\pm\boldsymbol{v} \in S^2$ consideriamo i cerchi ortogonali dati da $C_{\boldsymbol{v}} := \{\boldsymbol{x} \in S^2 : \langle \boldsymbol{x}, \boldsymbol{v} \rangle = 0\}$. (Tale $C_{\boldsymbol{v}}$ è l'equatore se consideriamo $\boldsymbol{v}$ come il polo nord della sfera.)

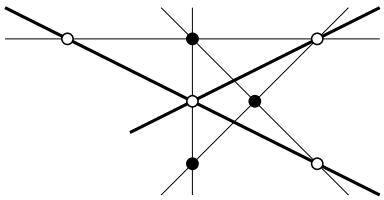
Allora il teorema di Sylvester-Gallai ci chiede di dimostrare:

Data una qualunque collezione di $n \geq 3$ cerchi massimi su S^2, non tutti passanti per uno stesso punto, vi è sempre un punto che giace esattamente su due dei cerchi massimi.

Ma la disposizione dei cerchi massimi fornisce un semplice grafo piano su S^2, i cui vertici sono i punti d'intersezione fra due di questi che dividono i cerchi massimi in archi. Tutti i gradi dei vertici sono pari e valgono almeno 4 per costruzione. Pertanto la parte (A) della proposizione assicura l'esistenza di un vertice di grado 4. È fatta! □

2. Rette monocromatiche

La dimostrazione seguente, analogo "colorato" del teorema di Sylvester-Gallai, si deve a Don Chakerian.



Teorema. *Per ogni configurazione finita di punti "bianchi" e "neri" nel piano, non tutti allineati, vi è sempre una retta "monocromatica", ovvero: una retta che contiene almeno due punti di uno stesso colore e nessuno dell'altro.*

■ **Dimostrazione.** Come per il problema di Sylvester-Gallai, trasferiamo il problema alla sfera unitaria e dualizziamolo lì. Dobbiamo così dimostrare:

Per ogni collezione finita di cerchi massimi "bianchi" e "neri" sulla sfera unitaria, non tutti passanti per un punto, vi è sempre un punto d'intersezione che giace o soltanto su cerchi massimi bianchi o soltanto su cerchi massimi neri.

Ora la risposta (positiva) è chiara dalla parte (C) della proposizione, dal momento che in ogni vertice in cui si intersechino cerchi massimi di colori diversi, abbiamo sempre almeno 4 angoli con cambiamenti di segni. □

3. Teorema di Pick

Il teorema di Pick del 1899 è un risultato splendido e sorprendente in sé, ma rappresenta anche una conseguenza "classica" della formula di Eulero. In seguito, definiremo *elementare* un poligono convesso $P \subseteq \mathbb{R}^2$ se i suoi vertici sono interi, ovvero giacciono sul *reticolo* $\mathbb{Z}^2$ e se non contiene alcun altro punto di tale reticolo.

Lemma. *Ogni triangolo elementare* $\Delta = \text{conv}\{\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2\} \subseteq \mathbb{R}^2$ *ha area* $A(\Delta) = \frac{1}{2}$.

■ **Dimostrazione.** Sia il parallelogramma P con angoli $\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_0$ che il reticolo $\mathbb{Z}^2$ sono simmetrici rispetto alla trasformazione

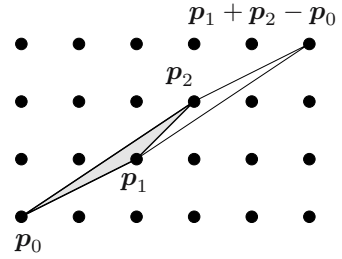
$$\sigma : \boldsymbol{x} \longmapsto \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{x},$$

Basi di un reticolo

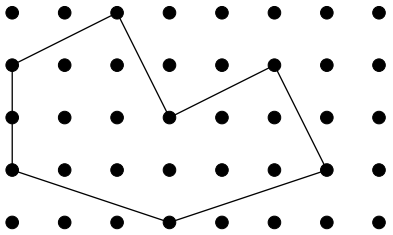
Una *base* di $\mathbb{Z}^2$ è una coppia di vettori linearmente indipendenti $\boldsymbol{e}_1, \boldsymbol{e}_2$ tale che

$$\mathbb{Z}^2 \;=\; \{\lambda_1 \boldsymbol{e}_1 + \lambda_2 \boldsymbol{e}_2 : \lambda_1, \lambda_2 \in \mathbb{Z}\}.$$

Siano $\boldsymbol{e}_1 = \binom{a}{b}$ e $\boldsymbol{e}_2 = \binom{c}{d}$, allora l'area del parallelogramma generato da $\boldsymbol{e}_1$ e $\boldsymbol{e}_2$ è data da $A(\boldsymbol{e}_1, \boldsymbol{e}_2) = |\det(\boldsymbol{e}_1, \boldsymbol{e}_2)| = |\det \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)|$. Se $\boldsymbol{f}_1 = \binom{r}{s}$ e $\boldsymbol{f}_2 = \binom{t}{u}$ è un'altra base, allora esiste una $\mathbb{Z}$-matrice Q invertibile con $\left(\begin{smallmatrix} r & t \\ s & u \end{smallmatrix}\right) = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right) Q$. Poiché $QQ^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ ed i determinanti sono numeri interi, ne segue che $|\det Q| = 1$ e pertanto $|\det(\boldsymbol{f}_1, \boldsymbol{f}_2)| = |\det(\boldsymbol{e}_1, \boldsymbol{e}_2)|$. Dunque tutti i parallelogrammi-base hanno la stessa area 1, poiché $A\big(\binom{1}{0}, \binom{0}{1}\big) = 1$.



che è la riflessione rispetto al centro del segmento congiungente $\boldsymbol{p}_1$ a $\boldsymbol{p}_2$. Allora il parallelogramma $P = \Delta \cup \sigma(\Delta)$ è anch'esso elementare ed i suoi traslati interi ricoprono il piano. Pertanto $\{\boldsymbol{p}_1 - \boldsymbol{p}_0, \boldsymbol{p}_2 - \boldsymbol{p}_0\}$ è una base per il reticolo $\mathbb{Z}^2$, ha come determinante ± 1, P è un parallelogramma di area 1 e Δ ha area $\frac{1}{2}$. Per una spiegazione di questi termini si veda il riquadro. □

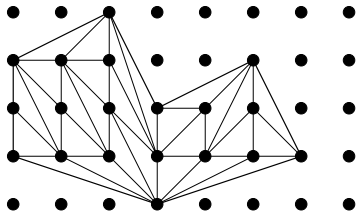


$n_{int} = 11,\ n_{bd} = 8,$ dunque $A = 14$

Teorema. *L'area di un qualsiasi poligono (non necessariamente convesso) $Q \subseteq \mathbb{R}^2$ con vertici interi è data da*

$$A(Q) \;=\; n_{int} + \frac{1}{2} n_{bd} - 1,$$

dove n_{int} e n_{bd} sono i numeri di punti interi rispettivamente nell'interno e sul bordo di Q.



■ **Dimostrazione.** Ogni poligono di questo tipo può essere triangolato usando tutti gli n_{int} punti del reticolo nell'interno e tutti gli n_{bd} punti del reticolo sul bordo di Q. Questo non è del tutto evidente, in particolare se non si richiede che Q sia convesso, ma l'argomentazione fornita al Capitolo 31 sul problema della galleria d'arte lo dimostra.

Interpretiamo ora la triangolazione come un grafo piano, che suddivide il piano in una faccia illimitata più $f - 1$ triangoli di area $\frac{1}{2}$, cosicché

$$A(Q) \;=\; \frac{1}{2}(f - 1).$$

Ogni triangolo ha tre lati e ognuno degli archi interni e_{int} limita due triangoli, mentre gli e_{bd} archi di bordo appaiono ciascuno in un singolo triangolo. Pertanto $3(f-1) = 2e_{int} + e_{bd}$ e dunque $f = 2(e - f) - e_{bd} + 3$. Inoltre, vi è lo stesso numero di archi di bordo che di vertici, $e_{bd} \;=\; n_{bd}$.

Questi due fatti insieme alla formula di Eulero danno

$$\begin{aligned} f &= 2(e-f) - e_{bd} + 3 \\ &= 2(n-2) - n_{bd} + 3 \;=\; 2n_{int} + n_{bd} - 1, \end{aligned}$$

e pertanto

$$A(Q) \;=\; \tfrac{1}{2}(f-1) \;=\; n_{int} + \tfrac{1}{2}n_{bd} - 1. \qquad \square$$

Bibliografia

[1] G. D. CHAKERIAN: *Sylvester's problem on collinear points and a relative,* Amer. Math. Monthly **77** (1970), 164-167.

[2] G. PICK: *Geometrisches zur Zahlenlehre,* Sitzungsberichte Lotos (Prag), Natur-med. Verein für Böhmen **19** (1899), 311-319.

[3] K. G. C. VON STAUDT: *Geometrie der Lage,* Verlag der Fr. Korn'schen Buchhandlung, Nürnberg 1847.

[4] N. E. STEENROD: *Solution 4065/Editorial Note,* Amer. Math. Monthly **51** (1944), 170-171.

Il teorema di rigidità di Cauchy

Capitolo 12

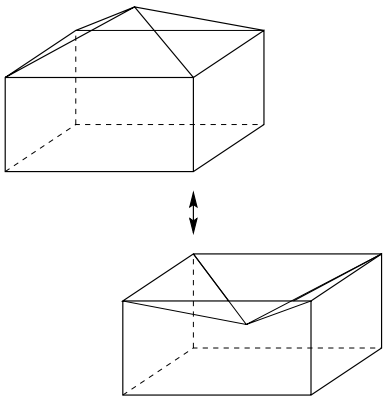
Un risultato celebre che dipende dalla formula di Eulero (e precisamente dalla parte (C) della proposizione nel capitolo precedente) è il teorema di rigidità di Cauchy per poliedri tridimensionali.

Per le nozioni di congruenza e di equivalenza combinatoria che sono usate in seguito facciamo riferimento all'appendice sui politopi e sui poliedri nel capitolo sul terzo problema di Hilbert (si veda a pagina 56).

Augustin Cauchy

> **Teorema.** *Se due poliedri convessi tridimensionali P e P' sono combinatorialmente equivalenti con le facce corrispondenti congruenti, allora anche gli angoli tra coppie corrispondenti di facce adiacenti sono uguali (e pertanto P è congruente a P').*

L'illustrazione a margine mostra due poliedri tridimensionali che sono combinatorialmente equivalenti e tali che le facce corrispondenti sono congruenti. Ma essi non sono congruenti e solo uno è convesso. Pertanto l'ipotesi di convessità è essenziale per il teorema di Cauchy!

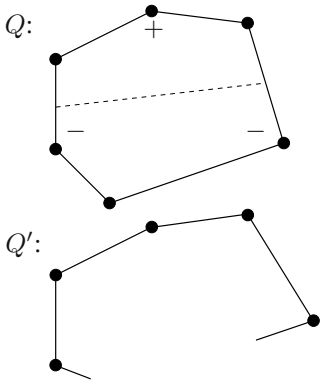


■ **Dimostrazione.** Quanto segue è essenzialmente la dimostrazione originale di Cauchy. Supponiamo che siano dati due poliedri convessi P e P' con facce congruenti. Coloriamo gli spigoli di P come segue: uno spigolo è nero (o "positivo") se il corrispondente angolo interno tra due facce adiacenti è più grande in P' che in P; è bianco (o "negativo") se l'angolo corrispondente è più piccolo in P' che in P.

Gli spigoli bianchi e neri di P formano insieme un grafo piano bicolore sulla superficie di P, che possiamo trasferire sulla sfera unitaria per proiezione radiale, supponendo che l'origine sia all'interno di P. Se P e P' hanno facce-angoli corrispondenti diversi, allora il grafo è non vuoto. Grazie alla parte (C) della proposizione nel capitolo precedente troviamo che vi è un vertice p adiacente ad almeno uno spigolo bianco o nero, tale che vi siano almeno due cambi tra spigoli neri e bianchi (in ordine ciclico).

Ora intersechiamo P con una piccola sfera S_ε (di raggio ε) con centro nel vertice p ed intersechiamo P' con una sfera S'_ε dello stesso raggio ε centrata nel corrispondente vertice p'. In S_ε ed S'_ε troviamo poligoni sferici convessi Q e Q' tali che gli archi corrispondenti hanno le stesse lunghezze, a causa della congruenza delle facce di P e P' e poiché abbiamo scelto lo stesso raggio ε.

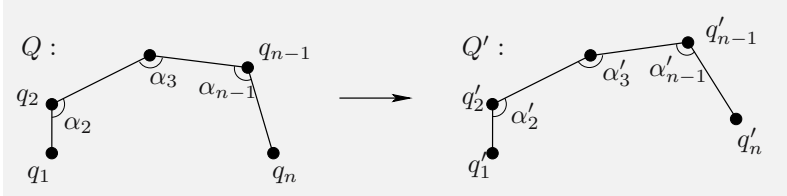
Contrassegnamo con $+$ gli angoli di Q per i quali gli angoli corrispondenti



in Q' sono più grandi, e con $-$ gli angoli i cui angoli corrispondenti in Q' sono più piccoli. Ovvero, quando ci spostiamo da Q a Q' gli angoli $+$ sono "aperti", gli angoli $-$ sono "chiusi", mentre tutte le lunghezze dei lati e gli angoli non contrassegnati restano costanti.

Dalla nostra scelta di p sappiamo che si verifica *qualche* segno $+$ o $-$ e che in ordine ciclico vi sono al massimo due cambi $+/-$. Se si verifica solo un tipo di segno, allora il lemma sottostante dà direttamente una contraddizione, poiché dice che uno spigolo deve cambiare la sua lunghezza. Se vi sono entrambi i tipi di segno, allora (poiché vi sono solo due cambi di segni) vi è una "retta di separazione" che congiunge i punti medi di due spigoli e separa tutti i segni $+$ da tutti i segni $-$. Nuovamente abbiamo una contraddizione dal lemma sottostante, dove la retta di separazione non può essere allo stesso tempo più lunga e più corta in Q' che in Q. □

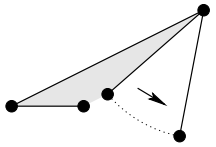
Lemma del braccio di Cauchy.
Se Q e Q' sono poligoni a n lati (piani o sferici), contrassegnati come nella figura,



tali che $\overline{q_i q_{i+1}} = \overline{q'_i q'_{i+1}}$ valga per le lunghezze degli spigoli corrispondenti per $1 \le i \le n-1$ e $\alpha_i \le \alpha'_i$ valga per le dimensioni degli angoli corrispondenti per $2 \le i \le n-1$, allora la lunghezza dello spigolo "mancante" soddisfa

$$\overline{q_1 q_n} \;\le\; \overline{q'_1 q'_n},$$

con uguaglianza se e solo si ha $\alpha_i = \alpha'_i$ per ogni i.



È interessante che la dimostrazione originale di Cauchy del lemma fosse falsa: un movimento continuo che apre angoli e mantiene fisse le lunghezze dei lati può distruggere la convessità — si veda la figura a lato! D'altro canto, sia il lemma che la sua dimostrazione data qui, tratta da una lettera di I. J. Schoenberg a S. K. Zaremba, sono validi per poligoni sia piani che sferici.

■ **Dimostrazione.** Procediamo per induzione su n. Il caso $n = 3$ è semplice: se in un triangolo aumentiamo l'angolo γ tra due lati di lunghezze fisse a e b, allora la lunghezza c del lato opposto aumenta a sua volta. Analiticamente, ciò segue dal teorema del coseno

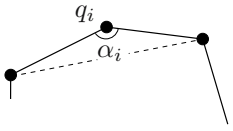
$$c^2 \;=\; a^2 + b^2 - 2ab\cos\gamma$$

nel caso piano e dal risultato analogo

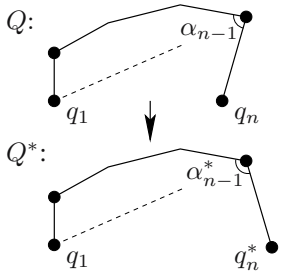
$$\cos c \;=\; \cos a \cos b + \sin a \sin b \cos\gamma$$

nella trigonometria sferica. Qui le lunghezze a, b, c si misurano sulla superficie di una sfera di raggio 1, e pertanto hanno valori nell'intervallo $[0, \pi]$.

Ora, sia $n \geq 4$. Se per ogni $i \in \{2, \ldots, n-1\}$ abbiamo $\alpha_i = \alpha'_i$, allora il vertice corrispondente può essere tagliato introducendo la diagonale da q_{i-1} a q_{i+1} (risp. da q'_{i-1} a q'_{i+1}), con $\overline{q_{i-1}q_{i+1}} = \overline{q'_{i-1}q'_{i+1}}$ e così terminiamo ragionando per induzione. Pertanto possiamo supporre $\alpha_i < \alpha'_i$ per $2 \leq i \leq n-1$.



Produciamo ora un nuovo poligono Q^* da Q sostituendo α_{n-1} con l'angolo più grande possibile $\alpha^*_{n-1} \leq \alpha'_{n-1}$ che mantenga Q^* convesso. Per fare ciò sostituiamo q_n con q^*_n, mantenendo tutti gli altri q_i, le lunghezze degli spigoli e gli angoli come in Q.



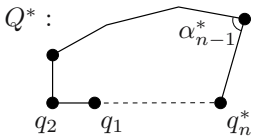
Se effettivamente possiamo scegliere $\alpha^*_{n-1} = \alpha'_{n-1}$ mantenendo Q^* convesso, allora abbiamo $\overline{q_1q_n} < \overline{q_1q^*_n} \leq \overline{q'_1q'_n}$, sfruttando il caso $n = 3$ per la prima fase e l'induzione come sopra per la seconda.

Altrimenti dopo un movimento non banale che dà

$$\overline{q_1q^*_n} \;>\; \overline{q_1q_n} \tag{1}$$

ci blocchiamo in una situazione in cui q_2, q_1 e q^*_n sono colineari, con

$$\overline{q_2q_1} + \overline{q_1q^*_n} \;=\; \overline{q_2q^*_n}. \tag{2}$$



Ora confrontiamo questo Q^* con Q' e troviamo

$$\overline{q_2q^*_n} \;\leq\; \overline{q'_2q'_n} \tag{3}$$

per induzione su n (ignorando il vertice q_1, risp. q'_1). Pertanto otteniamo

$$\overline{q'_1q'_n} \overset{(*)}{\geq} \overline{q'_2q'_n} - \overline{q'_1q'_2} \overset{(3)}{\geq} \overline{q_2q^*_n} - \overline{q_1q_2} \overset{(2)}{=} \overline{q_1q^*_n} \overset{(1)}{>} \overline{q_1q_n},$$

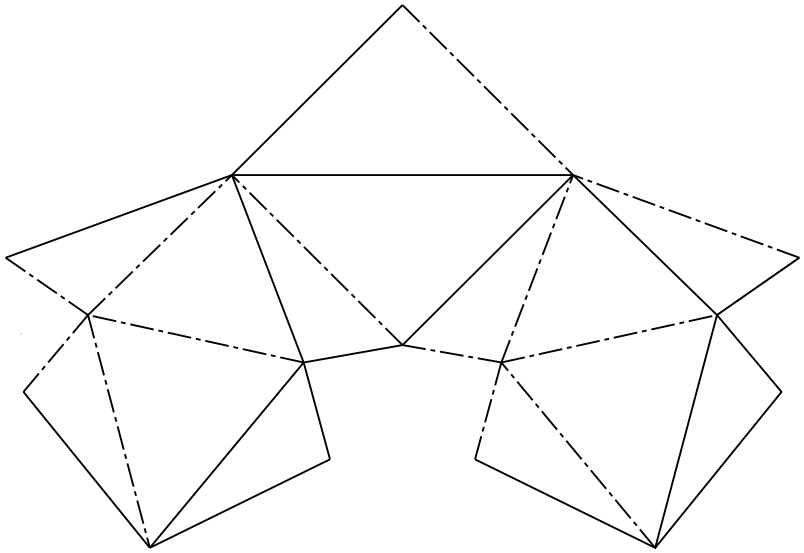
dove $(*)$ è semplicemente la disuguaglianza triangolare e tutte le altre relazioni sono già state derivate. □

Abbiamo visto un esempio che mostra che il teorema di Cauchy non è valido per poliedri *non convessi*. La caratteristica speciale di questo esempio è, ovviamente, che un "balzo" non continuo porta da un poliedro all'altro, mantenendo le facce congruenti mentre gli angoli diedri "saltano". Si può chiedere di più:

> *Potrebbe esistere, per un poliedro non convesso, una deformazione* continua *che mantenga le facce piane e congruenti?*

Fu congetturato che nessuna superficie triangolata, convessa o no, ammettesse un tale movimento. Fu dunque un'autentica sorpresa quando nel 1977 — più di 160 anni dopo l'opera di Cauchy — Robert Connelly presentò dei

controesempi: sfere chiuse triangolate in $\mathbb{R}^3$ (senza auto-interruzioni) flessibili, con un movimento continuo che mantiene costanti tutte le lunghezze degli spigoli, e pertanto mantiene le facce triangolari congruenti.



Uno splendido esempio di superficie flessibile costruito da Klaus Steffen: le linee tratteggiate rappresentano gli spigoli non convessi in questo modello di carta "intagliato". Si pieghino le linee piene in modo concavo e le linee tratteggiate in modo convesso. Gli spigoli nel modello hanno lunghezza di 5, 10, 11, 12 e 17 unità

La teoria della rigidità delle superfici ha ancora più sorprese in serbo: solo molto recentemente Idjad Sabitov è stato in grado di dimostrare che quando una qualunque di tali superfici flessibili si sposta, il *volume* che essa include deve essere costante. La sua dimostrazione è notevole anche per l'uso del macchinario algebrico di polinomi e determinanti (che va peraltro al di là degli scopi di questo libro).

Bibliografia

[1] A. CAUCHY: *Sur les polygones et les polyèdres, seconde mémoire,* J. École Polytechnique XVIe Cahier, Tome IX (1813), 87-98; Œuvres Complètes, IIe Série, Vol. 1, Paris 1905, 26-38.

[2] R. CONNELLY: *A counterexample to the rigidity conjecture for polyhedra,* Inst. Haut. Etud. Sci., Publ. Math. **47** (1978), 333-338.

[3] R. CONNELLY: *The rigidity of polyhedral surfaces,* Mathematics Magazine **52** (1979), 275-283.

[4] I. KH. SABITOV: *The volume as a metric invariant of polyhedra,* Discrete Comput. Geometry **20** (1998), 405-425.

[5] J. SCHOENBERG & S.K. ZAREMBA: *On Cauchy's lemma concerning convex polygons,* Canadian J. Math. **19** (1967), 1062-1071.

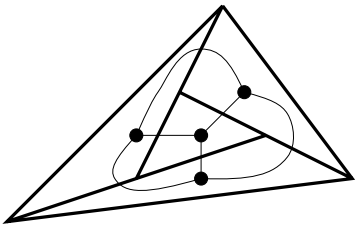
Simplessi contigui

Capitolo 13

Quanti simplessi d-dimensionali si possono posizionare in $\mathbb{R}^d$ in modo che si tocchino a due a due, ovvero, in modo che tutte le loro intersezioni a due a due siano $(d-1)$-dimensionali?

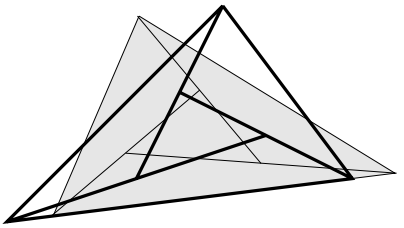
Questa è una domanda molto antica e naturale. Chiameremo $f(d)$ la risposta a questo problema, sapendo già che $f(1) = 2$, ovviamente. Per $d = 2$ la configurazione dei quattro triangoli a margine mostra che $f(2) \geq 4$. Non vi è alcuna configurazione simile con cinque triangoli, poiché la costruzione del grafo duale di tale configurazione, che nel nostro esempio con quattro triangoli dà un disegno planare di K_4, darebbe un'inclusione piana di K_5, il che è impossibile (si veda a pagina 73). Abbiamo pertanto

$$f(2) = 4.$$



$f(2) \geq 4$

In tre dimensioni possiamo facilmente osservare che $f(3) \geq 8$. Per questo usiamo la configurazione di otto triangoli presentata a destra. I quattro triangoli ombreggiati sono collegati con un punto x sotto il piano del disegno, che dà quattro tetraedri che toccano il piano da sotto. Analogamente, i quattro triangoli bianchi sono collegati con un punto y al di sopra del piano del disegno. Otteniamo così una configurazione di otto tetraedri che si toccano in $\mathbb{R}^3$, ovvero $f(3) \geq 8$.



$f(3) \geq 8$

Nel 1965 Baston scrisse un libro in cui dimostrava che $f(3) \leq 9$ e nel 1991 ci volle un altro libro perché Zaks stabilisse che

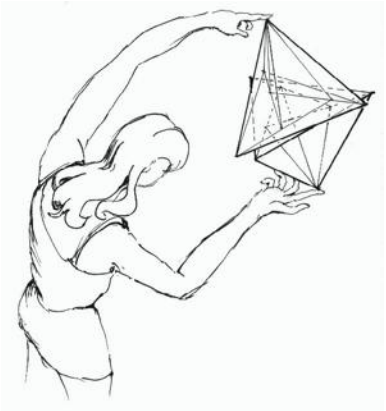
$$f(3) = 8.$$

Con $f(1) = 2, f(2) = 4$ e $f(3) = 8$ non ci vuole molta ispirazione per arrivare alla congettura seguente, formulata per la prima volta da Bagemihl nel 1956.

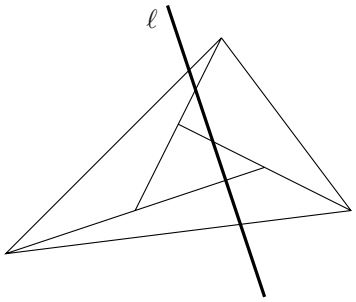
Congettura. *Il numero massimo di d-simplessi che si toccano a due a due in una configurazione in $\mathbb{R}^d$ è*

$$f(d) = 2^d.$$

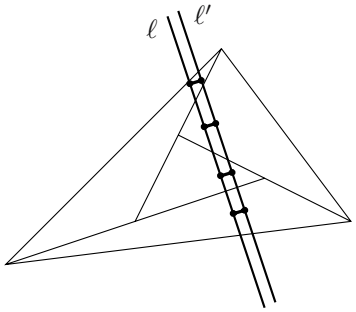
La minorazione $f(d) \geq 2^d$ è semplice da verificare "purché lo si faccia in modo opportuno". Questo si riduce ad un uso massiccio di trasformazioni affini di coordinate e ad un ragionamento per induzione sulla dimensione che fornisce il seguente risultato più forte, dovuto a Joseph Zaks [4].



"Simplessi che si toccano"



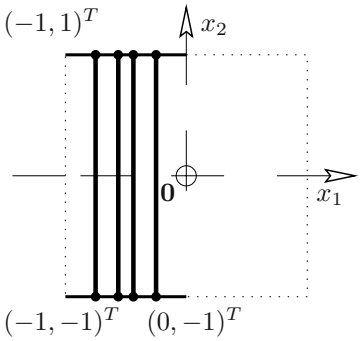
Teorema 1. *Per ogni $d \geq 2$, vi è una famiglia di 2^d d-simplessi che si toccano a due a due in $\mathbb{R}^d$ ed una retta trasversale che passa per l'interno di ognuno di essi.*



■ **Dimostrazione.** Per $d = 2$ la famiglia di quattro triangoli che abbiamo considerato ha una tale retta trasversale. Si consideri ora una qualunque configurazione d-dimensionale di simplessi che si toccano che abbia una retta trasversale ℓ. Qualunque retta parallela ℓ' nelle vicinanze è anch'essa trasversale. Se scegliamo ℓ' e ℓ parallele ed abbastanza vicine, allora ognuno dei simplessi contiene un intervallo ortogonale (il più breve) che colleghi le due rette. Solo una parte limitata delle rette ℓ e ℓ' è contenuta nei simplessi della configurazione e possiamo aggiungere due segmenti di collegamento al di fuori della configurazione, tali che il rettangolo generato dalle due rette esterne che essi collegano (ovvero il loro inviluppo convesso) contenga tutti gli altri segmenti di collegamento. Pertanto abbiamo posto una "scala" tale che ognuno dei simplessi della configurazione abbia uno dei gradini all'interno, mentre le quattro estremità della scala siano al di fuori della configurazione.

Il passo principale consiste ora nell'operare una trasformazione affine di coordinate fra $\mathbb{R}^d$ e $\mathbb{R}^d$, che trasforma il rettangolo generato dalla scala nel rettangolo (semi-quadrato) mostrato in figura, dato da

$$R^1 \;=\; \{(x_1, x_2, 0, \ldots, 0)^T : -1 \leq x_1 \leq 0; -1 \leq x_2 \leq 1\}.$$



Pertanto la configurazione di simplessi ottenuti Σ^1 in $\mathbb{R}^d$ che si toccano ha l'asse delle x_1 come retta trasversale ed è posto in modo tale che ognuno dei simplessi contenga un segmento

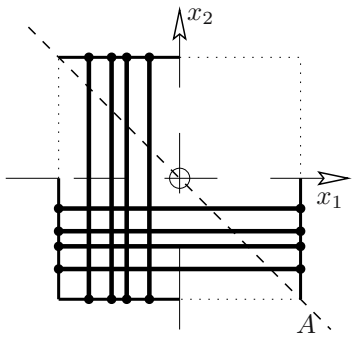
$$S^1(\alpha) \;=\; \{(\alpha, x_2, 0, \ldots, 0)^T : -1 \leq x_2 \leq 1\}$$

al suo interno (per qualche α dato, $-1 < \alpha < 0$), mentre l'origine $\mathbf{0}$ è al di fuori di tutti i simplessi.

Produciamo ora una seconda copia Σ^2 di questa configurazione riflettendo il primo nell'iperpiano dato da $x_1 = x_2$. Questa seconda configurazione ha l'asse x_2 come retta trasversale ed ogni simplesso contiene un segmento

$$S^2(\beta) \;=\; \{(x_1, \beta, 0, \ldots, 0)^T : -1 \leq x_1 \leq 1\}$$

al proprio interno, con $-1 < \beta < 0$. Ma ogni segmento $S^1(\alpha)$ interseca ogni segmento $S^2(\beta)$ e pertanto l'interno di ogni simplesso di Σ^1 interseca ogni simplesso di Σ^2 al proprio interno. Pertanto se aggiungiamo una $(d+1)$-esima coordinata, x_{d+1} e prendiamo come Σ



$$\{\text{conv}(P_i \cup \{-\boldsymbol{e}_{d+1}\}) : P_i \in \Sigma^1\} \;\cup\; \{\text{conv}(P_j \cup \{\boldsymbol{e}_{d+1}\}) : P_j \in \Sigma^2\},$$

otteniamo una configurazione di $(d+1)$-simplessi che si toccano in $\mathbb{R}^{d+1}$. Inoltre, l'antidiagonale

$$A \;=\; \{(x, -x, 0, \ldots, 0)^T : x \in \mathbb{R}\} \;\subseteq\; \mathbb{R}^d$$

interseca tutti i segmenti $S^1(\alpha)$ e $S^2(\beta)$. Inclinandola leggermente otteniamo una retta

$$L_\varepsilon \;=\; \{(x,-x,0,\ldots,0,\varepsilon x)^T : x\in\mathbb{R}\} \;\subseteq\; \mathbb{R}^{d+1},$$

che per ogni $\varepsilon > 0$ abbastanza piccolo interseca tutti i simplessi di Σ. Questo completa il nostro passo di induzione. □

Contrariamente a questa minorazione esponenziale, delle maggiorazioni strette sono più difficili da ottenere. Un'argomentazione ingenua per induzione (che considera separatamente tutti gli iperpiani delle facce in una configurazione di simplessi che si toccano) dà soltanto

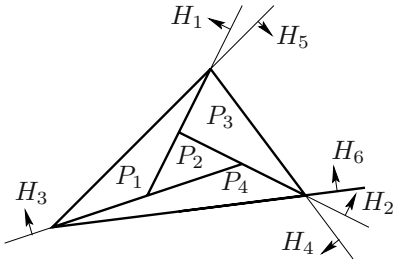
$$f(d) \;\leq\; \frac{2}{3}(d+1)!,$$

che è ben lontano dalla minorazione del Teorema 1. Tuttavia, Micha Perles trovò la seguente "magica" dimostrazione per ottenere una maggiorazione nettamente migliore.

Teorema 2. *Per ogni $d \geq 1$, abbiamo $f(d) < 2^{d+1}$.*

■ **Dimostrazione.** Data una configurazione di r d-simplessi che si toccano $P_1, P_2, \ldots, P_r$ in $\mathbb{R}^d$, si enumerino dapprima i diversi iperpiani $H_1, H_2, \ldots, H_s$ generati dalle facce dei P_i e per ognuno di essi si scelga arbitrariamente un lato positivo H_i^+ e si chiami l'altro lato H_i^-.

Ad esempio, per la configurazione bidimensionale di $r = 4$ triangoli illustrata a fianco troviamo $s = 6$ iperpiani (che sono rette se $d = 2$).



Da questi dati costruiamo la B-*matrice*, una matrice $(r \times s)$ con elementi in $\{+1, -1, 0\}$, come segue:

$$B_{ij} := \begin{cases} +1 & \text{se } P_i \text{ ha una faccia } H_j \text{ e } P_i \subseteq H_j^+, \\ -1 & \text{se } P_i \text{ ha una faccia } H_j \text{ e } P_i \subseteq H_j^-, \\ \;0 & \text{se } P_i \text{ non ha una faccia } H_j. \end{cases}$$

Per esempio, la configurazione bidimensionale a destra dà origine alla matrice

$$B = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ -1 & -1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 1 \end{pmatrix}.$$

Vale la pena di ricordare tre proprietà della B-matrice. Innanzitutto, dal momento che ogni d-simplesso ha $d+1$ facce, troviamo che ogni riga di B ha esattamente $d+1$ elementi non nulli e pertanto esattamente $s-(d+1)$ elementi nulli. In secondo luogo, stiamo trattando una configurazione di simplessi che si toccano a due a due e dunque per ogni coppia di righe troviamo una colonna in cui una riga ha un elemento $+1$, mentre l'elemento nell'altra riga è -1. Ovvero le righe sono diverse *anche se non consideriamo i loro elementi non nulli*. In terzo luogo, le righe di B "rappresentano"

i simplessi P_i, con

$$P_i \;=\; \bigcap_{j:B_{ij}=1} H_j^+ \;\cap\; \bigcap_{j:B_{ij}=-1} H_j^- . \qquad (*)$$

Ora deriviamo da B una nuova matrice C, in cui ogni riga di B sia sostituita con tutti i vettori riga che si possono generare da essa sostituendo tutti gli zeri con $+1$ oppure -1. Poiché ogni riga di B ha $s-d-1$ zeri e B ha r righe, la matrice C ha $2^{s-d-1}r$ righe.

Nel nostro esempio C è una matrice (32×6) che inizia con

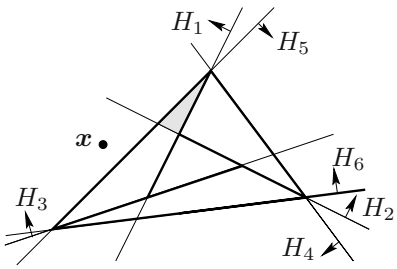
$$C = \left(\begin{array}{cccccc}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & -1 \\
1 & 1 & 1 & -1 & 1 & 1 \\
1 & 1 & 1 & -1 & 1 & -1 \\
1 & -1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & 1 & 1 & -1 \\
1 & -1 & 1 & -1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 & -1 \\
\hline
-1 & -1 & 1 & 1 & 1 & 1 \\
-1 & -1 & 1 & 1 & 1 & -1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array}\right),$$

dove le prime otto righe sono derivate dalla prima riga di B, le otto successive derivano dalla seconda riga di B, etc.

Il punto ora è che tutte le righe di C sono diverse: se due righe derivano dalla stessa riga di B, allora sono diverse poiché i loro zeri sono stati sostituiti diversamente; se derivano da diverse righe di B, allora differiscono indipendentemente da come siano stati sostituiti gli zeri. Ma le righe di C sono (± 1)-vettori di lunghezza s e vi sono solo 2^s vettori distinti di questo tipo. Pertanto dal momento che le righe di C sono diverse, C può avere al più 2^s righe, ovvero

$$2^{s-d-1}r \;\leq\; 2^s.$$

Tuttavia non tutti i (± 1)-vettori possibili appaiono in C, il che dà una disuguaglianza stretta $2^{s-d-1}r < 2^s$, e pertanto $r < 2^{d+1}$. Per verificarlo, notiamo che ogni riga di C rappresenta un'intersezione di semispazi — proprio come le righe di B prima, attraverso la formula $(*)$. Questa intersezione è un sottoinsieme del simplesso P_i, dato dalla riga corrispondente di B. Consideriamo un punto $\boldsymbol{x} \in \mathbb{R}^d$ che non giaccia su nessuno degli iperpiani H_j, né in alcuno dei simplessi P_i. Da questo $\boldsymbol{x}$ deriviamo un (± 1)-vettore che registra per ogni j se $\boldsymbol{x} \in H_j^+$ oppure $\boldsymbol{x} \in H_j^-$. Questo (± 1)-vettore non appare in C, perché l'intersezione del suo semispazio seguendo $(*)$ contiene $\boldsymbol{x}$ e pertanto non è contenuta in nessun simplesso P_i. □



La prima riga della matrice C rappresenta il triangolo ombreggiato, mentre la seconda riga corrisponde ad un'intersezione vuota dei semispazi. Il punto $\boldsymbol{x}$ porta al vettore

$$\begin{pmatrix} 1 & -1 & 1 & 1 & -1 & 1 \end{pmatrix}$$

che non appare nella C-matrice

Bibliografia

[1] F. BAGEMIHL: *A conjecture concerning neighboring tetrahedra,* Amer. Math. Monthly **63** (1956) 328-329.

[2] V. J. D. BASTON: *Some Properties of Polyhedra in Euclidean Space,* Pergamon Press, Oxford 1965.

[3] M. A. PERLES: *At most* 2^{d+1} *neighborly simplices in* E^d*,* Annals of Discrete Math. **20** (1984), 253-254.

[4] J. ZAKS: *Neighborly families of* 2^d *d-simplices in* E^d*,* Geometriae Dedicata **11** (1981), 279-296.

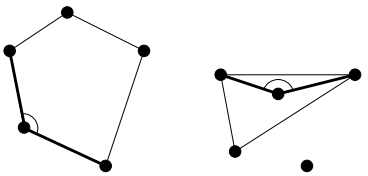
[5] J. ZAKS: *No Nine Neighborly Tetrahedra Exist,* Memoirs Amer. Math. Soc. No. 447, Vol. 91, 1991.

Ogni insieme grande di punti determina un angolo ottuso

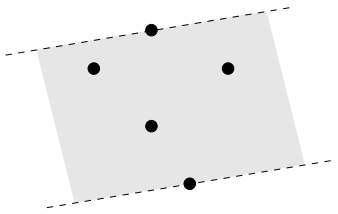
Capitolo 14

Intorno al 1950 Paul Erdős congetturò che ogni insieme di più di 2^d punti in $\mathbb{R}^d$ determina almeno un *angolo ottuso*, ovvero un angolo strettamente superiore a $\frac{\pi}{2}$. In altre parole, ogni insieme di punti in $\mathbb{R}^d$ che abbia solo angoli acuti (compresi gli angoli retti) ha come cardinalità massima 2^d. Questo problema fu posto come "domanda-premio" dalla Società Matematica Olandese, ma si ottennero soluzioni solo per $d = 2$ e per $d = 3$.

Per $d = 2$ il problema è semplice: i cinque punti possono determinare un pentagono convesso, che ha sempre un angolo ottuso (in effetti, ha almeno un angolo che misura non meno di $108°$). Altrimenti abbiamo un punto contenuto nell'inviluppo convesso di tre degli altri punti che formano un triangolo. Ma questo punto "vede" i tre lati del triangolo sotto tre angoli la cui somma è di $360°$, quindi uno degli angoli misura almeno $120°$. Il secondo caso comprende anche situazioni in cui abbiamo tre punti allineati e pertanto un angolo di $180°$.



Indipendentemente da questo, Victor Klee chiese qualche anno più tardi — ed Erdős diffuse la domanda — quanto potesse essere grande un insieme di punti in $\mathbb{R}^d$ per avere ancora la seguente "proprietà di antipodalità": per *ogni* coppia di punti vi è una striscia (limitata da due iperpiani paralleli) che contiene l'insieme di punti e che ha i due punti scelti su lati diversi del bordo.



In seguito, nel 1962, Ludwig Danzer e Branko Grünbaum risolsero entrambi i problemi in un solo colpo: inclusero entrambe le dimensioni massime in una catena di disuguaglianze, che inizia e termina con 2^d. Pertanto la risposta è 2^d sia per il problema di Erdős sia per quello di Klee.

Di seguito, consideriamo insiemi (finiti) di punti $S \subseteq \mathbb{R}^d$, i loro inviluppi convessi $\text{conv}(S)$ e i politopi generali convessi $Q \subseteq \mathbb{R}^d$. (Si veda l'appendice sui politopi a pagina 56 per i concetti di base). Supponiamo che questi insiemi abbiano la dimensione completa d, ovvero che non siano contenuti in un iperpiano. Due insiemi convessi *si toccano* (ovvero sono contigui) se hanno almeno un punto del bordo in comune, mentre i loro interni non si intersecano. Per ogni insieme $Q \subseteq \mathbb{R}^d$ ed ogni vettore $\boldsymbol{s} \in \mathbb{R}^d$ indichiamo con $Q + \boldsymbol{s}$ l'immagine di Q derivante dalla traslazione che sposta $\mathbf{0}$ in $\boldsymbol{s}$. Analogamente, $Q - \boldsymbol{s}$ è il traslato ottenuto tramite la trasformazione che sposta $\boldsymbol{s}$ nell'origine.

Non lasciatevi intimidire: questo capitolo è un'escursione nella geometria d-dimensionale, ma le argomentazioni che useremo non richiedono nessuna "intuizione in dimensione elevata", dal momento che possono tutte essere seguite, visualizzate (e pertanto *comprese*) in tre dimensioni o persino nel

piano. Pertanto le nostre figure illustreranno la dimostrazione per $d = 2$ (dove un "iperpiano" è semplicemente una retta) e potrete creare voi stessi le immagini per $d = 3$ (dove un "iperpiano" è un piano).

Teorema 1. *Per ogni d, si ha la seguente catena di disuguaglianze:*

$$
\begin{aligned}
2^d &\overset{(1)}{\leq} \max\left\{\#S \,\middle|\, S \subseteq \mathbb{R}^d,\ \sphericalangle(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k) \leq \tfrac{\pi}{2} \text{ per ogni } \{\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k\} \subseteq S\right\} \\
&\overset{(2)}{\leq} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tale che per ogni coppia di punti } \{\boldsymbol{s}_i, \boldsymbol{s}_j\} \subseteq \\ S \text{ vi è una striscia } \mathcal{S}(i,j) \text{ che contiene } S, \text{ con } \boldsymbol{s}_i \text{ e } \boldsymbol{s}_j \\ \text{giacenti negli iperpiani paralleli di bordo di } \mathcal{S}(i,j) \end{array}\right\} \\
&\overset{(3)}{=} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tale che i traslati } P - \boldsymbol{s}_i,\ \boldsymbol{s}_i \in S, \text{ dell' in-} \\ \text{viluppo convesso } P := \operatorname{conv}(S) \text{ si intersecano in un} \\ \text{punto comune, ma si toccano soltanto} \end{array}\right\} \\
&\overset{(4)}{\leq} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tale che i traslati } Q + \boldsymbol{s}_i \text{ di un politopo } d\text{-} \\ \text{dimensionale convesso } Q \subseteq \mathbb{R}^d \text{ si toccano due a due} \end{array}\right\} \\
&\overset{(5)}{=} \max\left\{\#S \,\middle|\, \begin{array}{l} S \subseteq \mathbb{R}^d \text{ tale che i traslati } Q^* + \boldsymbol{s}_i \text{ di un politopo} \\ d\text{-dimensionale convesso e centralmente simmetrico} \\ Q^* \subseteq \mathbb{R}^d \text{ si toccano due a due} \end{array}\right\} \\
&\overset{(6)}{\leq} 2^d.
\end{aligned}
$$

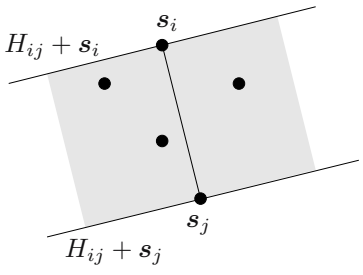
■ **Dimostrazione.** Abbiamo sei affermazioni (uguaglianze e disuguaglianze) da verificare. Diamoci da fare.

(1) Si prenda $S := \{0,1\}^d$ come insieme dei vertici del cubo unitario standard in $\mathbb{R}^d$ e si scelgano $\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k \in S$. Per simmetria possiamo supporre che $\boldsymbol{s}_j = \mathbf{0}$ sia il vettore nullo. Pertanto l'angolo si può calcolare con

$$\cos \sphericalangle(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{s}_k) = \frac{\langle \boldsymbol{s}_i, \boldsymbol{s}_k \rangle}{|\boldsymbol{s}_i||\boldsymbol{s}_k|}$$

che è chiaramente non negativo. Dunque S è un insieme privo di angoli ottusi tale che $|S| = 2^d$.

(2) Se S non contiene angoli ottusi, allora per ogni $\boldsymbol{s}_i, \boldsymbol{s}_j \in S$ possiamo definire $H_{ij} + \boldsymbol{s}_i$ e $H_{ij} + \boldsymbol{s}_j$ come gli iperpiani paralleli attraversanti $\boldsymbol{s}_i$ risp. $\boldsymbol{s}_j$ che siano ortogonali allo spigolo $[\boldsymbol{s}_i, \boldsymbol{s}_j]$.



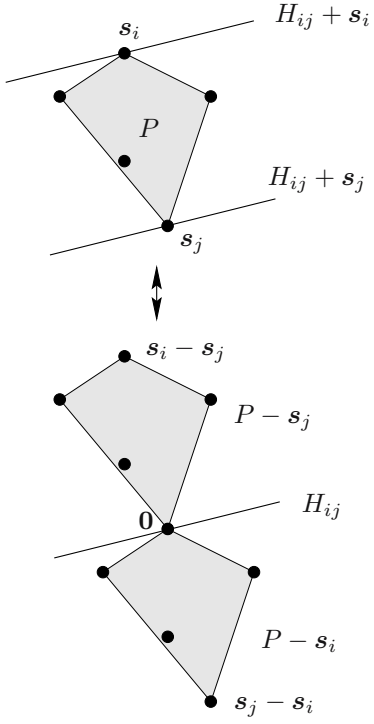
Qui $H_{ij} = \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{s}_i - \boldsymbol{s}_j \rangle = 0\}$ è l'iperpiano *passante per l'origine* ortogonale alla retta passante per $\boldsymbol{s}_i$ e $\boldsymbol{s}_j$ e $H_{ij} + \boldsymbol{s}_j = \{\boldsymbol{x} + \boldsymbol{s}_j : \boldsymbol{x} \in H_{ij}\}$ è il translato di H_{ij} che passa attraverso $\boldsymbol{s}_j$, etc. Pertanto la striscia fra $H_{ij} + \boldsymbol{s}_i$ e $H_{ij} + \boldsymbol{s}_j$ consiste, al di là di $\boldsymbol{s}_i$ e $\boldsymbol{s}_j$, esattamente in tutti i punti $\boldsymbol{x} \in \mathbb{R}^d$ tali che gli angoli $\sphericalangle(\boldsymbol{s}_i, \boldsymbol{s}_j, \boldsymbol{x})$ e $\sphericalangle(\boldsymbol{s}_j, \boldsymbol{s}_i, \boldsymbol{x})$ siano non ottusi. Dunque la striscia contiene tutto S.

(3) P è contenuto nel semispazio di $H_{ij} + \boldsymbol{s}_j$ che contiene $\boldsymbol{s}_i$ se e solo se $P - \boldsymbol{s}_j$ è contenuto nel semispazio di H_{ij} che contiene $\boldsymbol{s}_i - \boldsymbol{s}_j$: una proprietà che afferma che "un oggetto è contenuto in un semispazio" non viene meno se trasliamo sia l'oggetto che il semispazio della stessa quantità

(ovvero di $-\boldsymbol{s}_j$). Analogamente, P è contenuto nel semispazio di $H_{ij}+\boldsymbol{s}_i$ che contiene a sua volta $\boldsymbol{s}_j$ se e solo se $P-\boldsymbol{s}_i$ è contenuto nel semispazio di H_{ij} che contiene $\boldsymbol{s}_j-\boldsymbol{s}_i$.

Unendo le due affermazioni troviamo che il politopo P è contenuto nella striscia tra $H_{ij}+\boldsymbol{s}_i$ e $H_{ij}+\boldsymbol{s}_j$ se e solo se $P-\boldsymbol{s}_i$ e $P-\boldsymbol{s}_j$ giacciono in semispazi diversi rispetto all'iperpiano H_{ij}.

Questa corrispondenza è illustrata nello schizzo a margine.



Inoltre da $\boldsymbol{s}_i \in P = \mathrm{conv}(S)$ abbiamo che l'origine $\mathbf{0}$ è in tutti i traslati $P-\boldsymbol{s}_i$ ($\boldsymbol{s}_i \in S$). Pertanto vediamo che gli insiemi $P-\boldsymbol{s}_i$ si intersecano tutti in $\mathbf{0}$, ma sono soltanto contigui: i loro interni sono disgiunti a due a due, dal momento che essi giacciono su lati diversi degli iperpiani corrispondenti H_{ij}.

(4) Questo l'abbiamo gratis: la condizione che "i traslati devono toccarsi a due a due" è più debole della condizione che "si intersecano in un punto comune, ma si toccano soltanto". Analogamente, possiamo rilassare le condizioni permettendo che P sia un d-politopo convesso arbitrario in $\mathbb{R}^d$. Inoltre possiamo sostituire S con $-S$.

(5) Qui la parte "$\geq$" dell'uguaglianza è banale, ma non è quella che ci interessa. Dobbiamo iniziare con una configurazione $S \subseteq \mathbb{R}^d$ ed un d-politopo arbitrario $Q \subseteq \mathbb{R}^d$ tale che i traslati $Q+\boldsymbol{s}_i$ ($\boldsymbol{s}_i \in S$) si tocchino a due a due. Affermiamo che in questa situazione possiamo usare

$$Q^* := \left\{ \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{y}) \in \mathbb{R}^d : \boldsymbol{x},\boldsymbol{y} \in Q \right\}$$

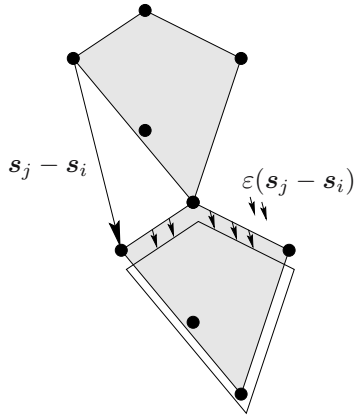
invece di Q. Ma questo non è difficile da verificare: innanzitutto, Q^* è d-dimensionale, convesso e a simmetria centrale. Si può controllare che Q^* sia un politopo (i suoi vertici sono della forma $\frac{1}{2}(\boldsymbol{q}_i-\boldsymbol{q}_j)$ per i vertici $\boldsymbol{q}_i,\boldsymbol{q}_j$ di Q), ma questo non è importante per noi.

Mostreremo ora che $Q+\boldsymbol{s}_i$ e $Q+\boldsymbol{s}_j$ si toccano *se e solo se* $Q^*+\boldsymbol{s}_i$ e $Q^*+\boldsymbol{s}_j$ si toccano. Per questo notiamo, seguendo le orme di Minkowski, che

$$\begin{aligned}
&(Q^*+\boldsymbol{s}_i)\cap(Q^*+\boldsymbol{s}_j)\neq\varnothing\\
&\quad\iff \exists\, \boldsymbol{q}_i',\boldsymbol{q}_i'',\boldsymbol{q}_j',\boldsymbol{q}_j'' \in Q : \tfrac{1}{2}(\boldsymbol{q}_i'-\boldsymbol{q}_i'')+\boldsymbol{s}_i = \tfrac{1}{2}(\boldsymbol{q}_j'-\boldsymbol{q}_j'')+\boldsymbol{s}_j\\
&\quad\iff \exists\, \boldsymbol{q}_i',\boldsymbol{q}_i'',\boldsymbol{q}_j',\boldsymbol{q}_j'' \in Q : \tfrac{1}{2}(\boldsymbol{q}_i'+\boldsymbol{q}_j'')+\boldsymbol{s}_i = \tfrac{1}{2}(\boldsymbol{q}_j'+\boldsymbol{q}_i'')+\boldsymbol{s}_j\\
&\quad\iff \exists\, \boldsymbol{q}_i,\boldsymbol{q}_j \in Q : \boldsymbol{q}_i+\boldsymbol{s}_i = \boldsymbol{q}_j+\boldsymbol{s}_j\\
&\quad\iff (Q+\boldsymbol{s}_i)\cap(Q+\boldsymbol{s}_j)\neq\varnothing,
\end{aligned}$$

dove nella terza (cruciale) equivalenza "$\iff$" sfruttiamo il fatto che ogni $\boldsymbol{q} \in Q$ si può scrivere come $\boldsymbol{q} = \frac{1}{2}(\boldsymbol{q}+\boldsymbol{q})$ per avere "$\Leftarrow$" e che Q è convesso e pertanto $\frac{1}{2}(\boldsymbol{q}_i'+\boldsymbol{q}_j'')$, $\frac{1}{2}(\boldsymbol{q}_j'+\boldsymbol{q}_i'') \in Q$ per vedere "$\Rightarrow$".

Pertanto il passaggio da Q a Q^* (noto come *simmetrizzazione di Minkowski*) conserva la proprietà che due traslati $Q+\boldsymbol{s}_i$ e $Q+\boldsymbol{s}_j$ si intersecano. Abbiamo dunque mostrato che per ogni insieme convesso Q, due traslati



$Q + s_i$ e $Q + s_j$ si intersecano se e solo se i traslati $Q^* + s_i$ e $Q^* + s_j$ si intersecano.

La caratterizzazione seguente mostra che la simmetrizzazione di Minkowski conserva anche la proprietà che due traslati si toccano:

> *$Q + s_i$ e $Q + s_j$ si toccano se e solo se si intersecano, mentre $Q + s_i$ e $Q + s_j + \varepsilon(s_j - s_i)$ non si intersecano per nessun $\varepsilon > 0$.*

(6) Supponiamo che $Q^* + s_i$ e $Q^* + s_j$ si tocchino. Per ogni punto di intersezione

$$x \in (Q^* + s_i) \cap (Q^* + s_j)$$

abbiamo

$$x - s_i \in Q^* \quad \text{e} \quad x - s_j \in Q^*,$$

dunque, poiché Q^* è a simmetria centrale,

$$s_i - x = -(x - s_i) \in Q^*,$$

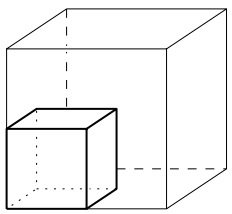
e pertanto, poiché Q^* è convesso,

$$\tfrac{1}{2}(s_i - s_j) = \tfrac{1}{2}\left((x - s_j) + (s_i - x)\right) \in Q^*.$$

Concludiamo che $\frac{1}{2}(s_i + s_j)$ è contenuto in $Q^* + s_j$ per ogni i. Di conseguenza, essendo $P := \text{conv}(S)$ abbiamo

$$P_j := \tfrac{1}{2}(P + s_j) = \text{conv}\left\{\tfrac{1}{2}(s_i + s_j) : s_i \in S\right\} \subseteq Q^* + s_j,$$

che implica che gli insiemi $P_j = \frac{1}{2}(P + s_j)$ possono solo essere contigui. Infine, gli insiemi P_j sono contenuti in P, poiché dal momento che tutti i punti s_i, s_j e $\frac{1}{2}(s_i + s_j)$ sono contenuti in P, P è convesso. Ma i P_j sono semplicemente traslati ridotti di P, contenuti in P. Il fattore di normalizzazione è $\frac{1}{2}$, il che implica che



Fattore di normalizzazione $\frac{1}{2}$, $\text{vol}(P_j) = \frac{1}{8}\text{vol}(P)$

$$\text{vol}(P_j) \;=\; \frac{1}{2^d}\text{vol}(P),$$

poiché stiamo trattando insiemi d-dimensionali. Ciò significa che al massimo 2^d insiemi P_j rientrano in P e pertanto $|S| \leq 2^d$.
Questo completa la nostra dimostrazione: la catena di disuguaglianze è chiusa. □

...ma non finisce qui. Danzer e Grünbaum posero il seguente naturale interrogativo:

> *Cosa accade se si richiede che tutti gli angoli siano **acuti** invece che semplicemente non ottusi ovvero se si escludono gli angoli retti?*

Essi costruirono configurazioni di $2d-1$ punti in $\mathbb{R}^d$ con solo angoli acuti, congetturando che questo potesse essere il migliore risultato possibile. Grünbaum dimostrò che ciò è in effetti vero per $d \leq 3$. Ma ventun anni più tardi, nel 1983, Paul Erdős e Zoltan Füredi mostrarono che la congettura è falsa — e notevolmente, se la dimensione è alta! La loro dimostrazione è un grande esempio del potere delle argomentazioni probabilistiche; si veda al Capitolo 35 per una introduzione al "metodo probabilistico". La nostra versione della dimostrazione usa un leggero miglioramento nella scelta dei parametri dovuto al nostro lettore David Bevan.

Teorema 2. *Per ogni $d \geq 2$, vi è un insieme $S \subseteq \{0,1\}^d$ di $2\lfloor \frac{\sqrt{6}}{9}\big(\frac{2}{\sqrt{3}}\big)^d \rfloor$ punti in $\mathbb{R}^d$ (vertici del cubo unitario d-dimensionale) che determinano solo angoli acuti.*
In particolare, in dimensione $d = 34$ vi è un insieme di $72 > 2{\cdot}34-1$ punti con solo angoli acuti.

■ **Dimostrazione.** Si ponga $m := \lfloor \frac{\sqrt{6}}{9}\big(\frac{2}{\sqrt{3}}\big)^d \rfloor$ e si prendano $3m$ vettori

$$\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(3m) \;\in\; \{0,1\}^d$$

scegliendo tutte le loro coordinate indipendentemente ed in modo casuale fra 0 o 1, con probabilità $\frac{1}{2}$ per ogni alternativa. (Si può lanciare una moneta perfetta $3md$ volte; tuttavia, se d è grande questa operazione può diventare rapidamente noiosa).

Abbiamo visto più sopra che tutti gli angoli determinati da $0/1$-vettori sono non ottusi. Tre vettori $\boldsymbol{x}(i), \boldsymbol{x}(j), \boldsymbol{x}(k)$ determinano un angolo retto con apice $\boldsymbol{x}(j)$ se e solo se il prodotto scalare $\langle \boldsymbol{x}(i) - \boldsymbol{x}(j), \boldsymbol{x}(k) - \boldsymbol{x}(j) \rangle$ si annulla ovvero se abbiamo

$$x(i)_\ell - x(j)_\ell = 0 \quad \text{oppure} \quad x(k)_\ell - x(j)_\ell = 0 \quad \text{per ogni coordinata } \ell.$$

Se ciò accade chiamiamo (i,j,k) una *cattiva tripletta*. (Se $\boldsymbol{x}(i) = \boldsymbol{x}(j)$ o $\boldsymbol{x}(j) = \boldsymbol{x}(k)$, allora l'angolo non è definito, ma non solo, anche la tripletta (i,j,k) è certamente cattiva.)

La probabilità che una specifica tripletta sia cattiva è esattamente $\big(\frac{3}{4}\big)^d$: in effetti, essa sarà buona se e solo se, per una delle d coordinate ℓ, abbiamo

$$\begin{array}{ll} \text{o} & x(i)_\ell = x(k)_\ell = 0, \quad x(j)_\ell = 1, \\ \text{o} & x(i)_\ell = x(k)_\ell = 1, \quad x(j)_\ell = 0. \end{array}$$

Questo ci lascia con sei cattive opzioni su otto ugualmente probabili ed una tripletta sarà cattiva se e solo se una delle cattive opzioni (con probabilità $\frac{3}{4}$) si verifica per ognuna delle d coordinate.

Il numero di triplette che dobbiamo considerare è $3\binom{3m}{3}$, poiché vi sono $\binom{3m}{3}$ insiemi di tre vettori e per ognuno di essi vi sono tre scelte possibili per l'apice. Ovviamente le probabilità che le varie triplette siano cattive non sono indipendenti: ma la *linearità del valore atteso* (che si ottiene facendo

la media di tutte le selezioni possibili; si veda all'appendice) dà come risultato che il numero *previsto* di cattive triplette è esattamente $3\binom{3m}{3}\left(\frac{3}{4}\right)^d$. Ciò significa — e questo è il punto in cui il metodo probabilistico mostra la sua forza — che vi è *una qualche* scelta dei $3m$ vettori tale che vi siano almeno $3\binom{3m}{3}\left(\frac{3}{4}\right)^d$ cattive triplette, dove

$$3\tbinom{3m}{3}\left(\tfrac{3}{4}\right)^d \;<\; 3\tfrac{(3m)^3}{6}\left(\tfrac{3}{4}\right)^d \;=\; m^3\left(\tfrac{9}{\sqrt{6}}\right)^2\left(\tfrac{3}{4}\right)^d \;\leq\; m,$$

in base alla scelta di m.

Ma se non vi sono più di m cattive triplette, allora possiamo rimuovere m dei $3m$ vettori $\boldsymbol{x}(i)$ in modo tale che i restanti $2m$ vettori non contengano cattive triplette ovvero che determinino solo angoli acuti. □

La "costruzione probabilistica" di un insieme grande di punti 0 e 1 senza angoli retti si può implementare facilmente, usando un generatore di numeri aleatori per "lanciare la monetina". David Bevan ha costruito in questo modo un insieme di 31 punti in dimensione $d = 15$ che determinano solo angoli acuti.

Appendice: tre strumenti dal calcolo delle probabilità

Riuniamo qui tre strumenti basilari dalla teoria discreta della probabilità che torneranno in gioco diverse volte: variabili aleatorie, linearità del valore atteso e disuguaglianza di Markov.

Sia (Ω, p) uno *spazio di probabilità* finito, ovvero Ω è un insieme finito e $p = \text{Prob}$ è una trasformazione di Ω nell'intervallo $[0,1]$ con $\sum_{\omega\in\Omega} p(\omega) = 1$. Una *variabile aleatoria* X su Ω è una trasformazione $X : \Omega \longrightarrow \mathbb{R}$. Definiamo uno spazio di probabilità sull'insieme delle immagini $X(\Omega)$ ponendo $p(X = x) := \sum_{X(\omega)=x} p(\omega)$. Un semplice esempio è un dado non truccato (tutti i $p(\omega) = \frac{1}{6}$) con $X =$ "il numero che esce sulla faccia superiore del dado lanciato".

Il *valore atteso* EX di X è la media prevedibile, ovvero

$$EX \;=\; \sum_{\omega\in\Omega} p(\omega)X(\omega).$$

Supponiamo ora che X e Y siano due variabili aleatorie su Ω; allora la somma $X + Y$ è di nuovo una variabile aleatoria ed otteniamo

$$\begin{aligned} E(X+Y) \;&=\; \sum_{\omega} p(\omega)(X(\omega)+Y(\omega)) \\ &=\; \sum_{\omega} p(\omega)X(\omega) + \sum_{\omega} p(\omega)Y(\omega) \;=\; EX + EY. \end{aligned}$$

Chiaramente, questo si può estendere ad una qualunque combinazione lineare di variabili aleatorie — questo è ciò che si definisce *linearità del valore atteso*. Si noti che non vi è alcun bisogno che le variabili aleatorie siano "indipendenti" in alcun senso!

Il nostro terzo strumento riguarda variabili aleatorie X che prendono solo valori non negativi, abbreviate con $X \geq 0$. Sia

$$\text{Prob}(X \geq a) \;=\; \sum_{\omega : X(\omega) \geq a} p(\omega)$$

la probabilità che X sia almeno grande quanto un $a > 0$. Allora

$$EX \;=\; \sum_{\omega : X(\omega) \geq a} p(\omega)X(\omega) \;+\; \sum_{\omega : X(\omega) < a} p(\omega)X(\omega) \;\geq\; a \sum_{\omega : X(\omega) \geq a} p(\omega),$$

ed abbiamo dimostrato la *disuguaglianza di Markov*

$$\text{Prob}(X \geq a) \;\leq\; \frac{EX}{a}.$$

Bibliografia

[1] L. DANZER & B. GRÜNBAUM: *Über zwei Probleme bezüglich konvexer Körper von P. Erdös und von V. L. Klee,* Math. Zeitschrift **79** (1962), 95-99.

[2] P. ERDŐS & Z. FÜREDI: *The greatest angle among n points in the d-dimensional Euclidean space,* Annals of Discrete Math. **17** (1983), 275-283.

[3] H. MINKOWSKI: *Dichteste gitterförmige Lagerung kongruenter Körper,* Nachrichten Ges. Wiss. Göttingen, Math.-Phys. Klasse 1904, 311-355.

La congettura di Borsuk

Capitolo 15

La pubblicazione di Karol Borsuk "Tre teoremi sulla sfera euclidea n-dimensionale" del 1933 è celebre perché contiene un risultato importante (congetturato da Stanisław Ulam) ora noto come il teorema di Borsuk-Ulam:

> *Ogni trasformazione continua $f : S^d \to \mathbb{R}^d$ trasforma due punti agli antipodi della sfera S^d nello stesso punto in $\mathbb{R}^d$.*

La stessa pubblicazione è celebre anche per un problema posto alla fine, che divenne noto come congettura di Borsuk:

> *È possibile partizionare un qualunque insieme $S \subseteq \mathbb{R}^d$ di diametro limitato diam$(S) > 0$ in al più $d+1$ insiemi di diametro inferiore?*

Karol Borsuk

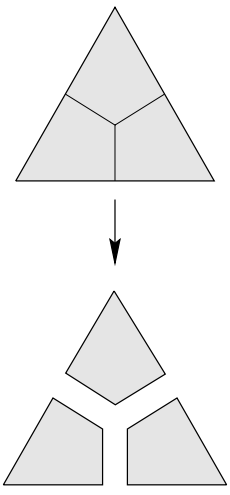
Il limite di $d + 1$ è il migliore possibile: se S è un simplesso regolare d-dimensionale, o semplicemente l'insieme dei suoi $d + 1$ vertici, allora nessuna parte di una partizione che riduca il diametro può contenere più di uno dei vertici del simplesso. Se $f(d)$ indica il più piccolo numero tale che ogni insieme limitato $S \subseteq \mathbb{R}^d$ abbia una partizione in $f(d)$ parti che riduca il diametro, allora l'esempio di un simplesso regolare stabilisce che $f(d) \geq d + 1$.

La congettura di Borsuk fu dimostrata per il caso in cui S è una sfera (dallo stesso Borsuk), per corpi lisci S (mediante il teorema di Borsuk-Ulam), per $d \leq 3$, ... ma la congettura generale rimase aperta. La miglior maggiorazione disponibile per $f(d)$ fu stabilita da Oded Schramm, il quale dimostrò che

$$f(d) \leq (1.23)^d$$

per d grande abbastanza. Questo limite sembra piuttosto debole in confronto alla congettura "$f(d) = d + 1$", ma parse improvvisamente ragionevole quando Jeff Kahn e Gil Kalai invalidarono con grande scalpore la congettura di Borsuk nel 1993. Sessant'anni dopo la pubblicazione di Borsuk, Kahn e Kalai dimostrarono che, per d abbastanza grandi, $f(d) \geq (1.2)^{\sqrt{d}}$.

Una versione da *Book Proof* della dimostrazione di Kahn-Kalai fu fornita da A. Nilli: breve ed autosufficiente, dà un controesempio esplicito della congettura di Borsuk in dimensione $d = 946$. Presentiamo qui una versione modificata della dimostrazione, ad opera di Andrei M. Raigorodskii e di Bernulf Weißbach, che riduce la dimensione a $d = 561$ e persino a $d = 560$.



Ogni d-simplesso *può* essere diviso in $d + 1$ parti, ognuna delle quali di diametro inferiore

A. Nilli

L'attuale "record" è $d = 298$, raggiunto da Aicke Hinrichs e Christian Richter nel 2002.

Teorema. *Sia $q = p^m$ una potenza prima, $n := 4q-2$, e $d := \binom{n}{2} = (2q-1)(4q-3)$. Allora esiste un insieme $S \subseteq \{+1,-1\}^d$ di 2^{n-2} punti in $\mathbb{R}^d$ tale che ogni partizione di S, le cui parti abbiano un diametro inferiore rispetto a S, abbia almeno*

$$\frac{2^{n-2}}{\sum_{i=0}^{q-2}\binom{n-1}{i}}$$

parti. Per $q = 9$ questo implica che la congettura di Borsuk è falsa in dimensione $d = 561$. Inoltre, $f(d) > (1.2)^{\sqrt{d}}$ vale per tutti i valori di d abbastanza grandi.

■ **Dimostrazione.** La costruzione dell'insieme S si svolge in quattro fasi.

(1) Sia q una potenza prima, si ponga $n = 4q-2$ e sia

$$Q \;:=\; \Big\{\boldsymbol{x} \in \{+1,-1\}^n : x_1 = 1,\ \#\{i : x_i = -1\} \text{ è pari}\Big\}.$$

Questo Q è un insieme di 2^{n-2} vettori in $\mathbb{R}^n$. Vedremo che $\langle \boldsymbol{x}, \boldsymbol{y}\rangle \equiv 2 \pmod 4$ vale per tutti i vettori $\boldsymbol{x}, \boldsymbol{y} \in Q$; diremo che $\boldsymbol{x}, \boldsymbol{y}$ sono *quasi ortogonali* se $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = 2$; dimostreremo che ogni sottoinsieme $Q' \subseteq Q$ che non contenga vettori quasi ortogonali deve essere piccolo: $|Q'| \le \sum_{i=0}^{q-2}\binom{n-1}{i}$.

(2) Da Q costruiamo l'insieme

$$R \;:=\; \{\boldsymbol{x}\boldsymbol{x}^T : \boldsymbol{x} \in Q\}$$

$$\boldsymbol{x} \;=\; \begin{pmatrix} 1\\ -1\\ -1\\ 1\\ -1\end{pmatrix} \quad\Longrightarrow$$

$$\boldsymbol{x}^T \;=\; \begin{pmatrix} 1 & -1 & -1 & 1 & -1\end{pmatrix}$$

$$\boldsymbol{x}\boldsymbol{x}^T = \begin{pmatrix} 1 & -1 & -1 & 1 & -1\\ -1 & 1 & 1 & -1 & 1\\ -1 & 1 & 1 & -1 & 1\\ 1 & -1 & -1 & 1 & -1\\ -1 & 1 & 1 & -1 & 1\end{pmatrix}$$

Vettori, matrici e prodotti scalari

Nella nostra notazione tutti i vettori $\boldsymbol{x}, \boldsymbol{y}, \ldots$ sono vettori-colonna; i vettori trasposti $\boldsymbol{x}^T, \boldsymbol{y}^T, \ldots$ sono pertanto vettori-riga. Il prodotto matriciale $\boldsymbol{x}\boldsymbol{x}^T$ è una matrice di rango 1, con $(\boldsymbol{x}\boldsymbol{x}^T)_{ij} = x_i x_j$.
Se $\boldsymbol{x}, \boldsymbol{y}$ sono vettori-colonna, allora il loro *prodotto scalare* è

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle \;=\; \sum_i x_i y_i \;=\; \boldsymbol{x}^T\boldsymbol{y}.$$

Avremo anche bisogno dei prodotti scalari per le matrici $X, Y \in \mathbb{R}^{n\times n}$ che si possono interpretare come vettori di lunghezza n^2 e pertanto il loro prodotto scalare è

$$\langle X, Y\rangle \;:=\; \sum_{i,j} x_{ij} y_{ij}.$$

delle 2^{n-2} matrici simmetriche $(n \times n)$ di rango 1 che interpretiamo come vettori con n^2 componenti, $R \subseteq \mathbb{R}^{n^2}$. Mostreremo che vi sono solo angoli acuti fra questi vettori: essi danno prodotti scalari positivi che valgono almeno 4. Inoltre, se $R' \subseteq R$ non contiene alcuna coppia di vettori con un prodotto scalare minimo di 4, allora $|R'|$ è "piccolo": $|R'| \leq \sum_{i=0}^{q-2} \binom{n-1}{i}$.

(3) Da R otteniamo l'insieme dei punti in $\mathbb{R}^{\binom{n}{2}}$ le cui coordinate sono gli elementi sotto la diagonale delle matrici corrispondenti:

$$S \;:=\; \{(\boldsymbol{x}\boldsymbol{x}^T)_{i>j} : \boldsymbol{x}\boldsymbol{x}^T \in R\}.$$

Nuovamente, S consiste in 2^{n-2} punti. La distanza massima tra questi punti si ottiene precisamente per i vettori quasi ortogonali $\boldsymbol{x}, \boldsymbol{y} \in Q$. Concludiamo che un sottoinsieme $S' \subseteq S$ di diametro inferiore a S deve essere piccolo: $|S'| \leq \sum_{i=0}^{q-2} \binom{n-1}{i}$.

(4) Stime: da **(3)** vediamo che servono almeno

$$g(q) \;:=\; \frac{2^{4q-4}}{\sum_{i=0}^{q-2} \binom{4q-3}{i}}$$

parti in ogni partizione che riduce il diametro di S. Pertanto

$$f(d) \;\geq\; \max\{g(q), d+1\} \qquad \text{per } d \;=\; (2q-1)(4q-3).$$

Dunque ogniqualvolta abbiamo $g(q) > (2q-1)(4q-3)+1$, troviamo un controesempio della congettura di Borsuk in dimensione $d = (2q-1)(4q-3)$.

Calcoleremo più avanti che $g(9) > 562$, il che fornisce il controesempio in dimensione $d = 561$ e che

$$g(q) \;>\; \frac{e}{64\,q^2}\left(\frac{27}{16}\right)^q,$$

il che dà la stima asintotica $f(d) > (1.2)^{\sqrt{d}}$ per d grande abbastanza.

Dettagli di (1): Inizieremo con alcune innocue considerazioni sulla divisibilità.

Lemma. *La funzione $P(z) := \binom{z-2}{q-2}$ è un polinomio di grado $q-2$. Essa dà valori interi per tutti gli interi z. L'intero $P(z)$ è divisibile per p se e solo se z non è congruente con 0 o 1 modulo q.*

■ **Dimostrazione.** Scriviamo il coefficiente binomiale come

$$P(z) \;=\; \binom{z-2}{q-2} \;=\; \frac{(z-2)(z-3)\cdot\ldots\cdot(z-q+1)}{(q-2)(q-3)\cdot\;\ldots\;\ldots\;\cdot 2\cdot 1} \qquad (*)$$

e confrontiamo il numero di p-fattori nel denominatore e nel numeratore. Il denominatore ha lo stesso numero di p-fattori di $(q-2)!$ oppure di $(q-1)!$ poiché $q-1$ non è divisible per p. In effetti, secondo l'affermazione nella

Proposizione. *Se $a \equiv b \not\equiv 0 \pmod q$, allora a e b hanno lo stesso numero di p-fattori.*

■ **Dimostrazione.** Abbiamo $a = b + sp^m$, dove b non è divisibile da $p^m = q$. Così ogni potenza p^k che divide b soddisfa $k < m$ e pertanto divide anche a. L'affermazione è simmetrica in a e b. □

pagina precedente abbiamo un intero con lo stesso numero di p-fattori se prendiamo *ogni* prodotto di $q-1$ interi, uno per ogni classe residua non nulla modulo q.

Ora se z è congruente a 0 o $1 \pmod q$, allora lo è anche il numeratore: tutti i fattori nel prodotto provengono da classi residue diverse e le sole classi che non si incontrano sono le classi nulle (i multipli di q); la classe di -1 o di $+1$ è divisibile per p, ma non quella di $+1$ o di -1. Pertanto il denominatore ed il numeratore hanno lo stesso numero di p-fattori, dunque il quoziente non è divisibile per p.

D'altro canto, se $z \not\equiv 0, 1 \pmod q$, allora il numeratore di $(*)$ contiene un fattore divisibile per $q = p^m$. Allo stesso tempo, il prodotto non ha fattori da due classi residue adiacenti non nulle: una di esse rappresenta i numeri che non hanno p-fattori, mentre l'altra ha meno p-fattori che $q = p^m$. Pertanto vi sono più p-fattori nel numeratore che nel denominatore ed il quoziente è divisibile per p. □

Ora consideriamo un insieme arbitrario $Q' \subseteq Q$ che non contenga vettori quasi ortogonali. Vogliamo stabilire che Q' deve essere piccolo.

Proposizione 1. *Se $\boldsymbol{x}, \boldsymbol{y}$ sono vettori distinti presi in Q, allora $\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)$ è un intero nell'intervallo*

$$-(q-2) \;\le\; \tfrac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2) \;\le\; q-1.$$

Sia $\boldsymbol{x}$ sia $\boldsymbol{y}$ hanno un numero pari di componenti uguali a (-1), dunque il numero di componenti in cui $\boldsymbol{x}$ ed $\boldsymbol{y}$ differiscono è anch'esso pari. Pertanto

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle \;=\; (4q-2) \;-\; 2\#\{i : x_i \neq y_i\} \;\equiv\; -2 \pmod 4$$

per ogni $\boldsymbol{x}, \boldsymbol{y} \in Q$, ovvero $\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)$ è un intero.
Da $\boldsymbol{x}, \boldsymbol{y} \in \{+1, -1\}^{4q-2}$ vediamo che $-(4q-2) \le \langle \boldsymbol{x}, \boldsymbol{y}\rangle \le 4q-2$, ovvero $-(q-1) \le \frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2) \le q$. La minorazione non vale mai con uguaglianza, dal momento che $x_1 = y_1 = 1$ implica che $\boldsymbol{x} \neq -\boldsymbol{y}$. La maggiorazione vale con uguaglianza solo se $\boldsymbol{x} = \boldsymbol{y}$.

Proposizione 2. *Per ogni $\boldsymbol{y} \in Q'$, il polinomio in n variabili $x_1, \ldots, x_n$ di grado $q-2$ dato da*

$$F_{\boldsymbol{y}}(\boldsymbol{x}) \;:=\; P\big(\tfrac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)\big) \;=\; \binom{\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2) - 2}{q-2}$$

è divisibile per p per ogni $\boldsymbol{x} \in Q' \setminus \{\boldsymbol{y}\}$, ma non per $\boldsymbol{x} = \boldsymbol{y}$.

La rappresentazione mediante un coefficiente binomiale mostra che $F_{\boldsymbol{y}}(\boldsymbol{x})$ è un polinomio a valori interi. Per $\boldsymbol{x} = \boldsymbol{y}$ abbiamo $F_{\boldsymbol{y}}(\boldsymbol{y}) = 1$. Per $\boldsymbol{x} \neq \boldsymbol{y}$ dal Lemma segue che $F_{\boldsymbol{y}}(\boldsymbol{x})$ non è divisibile per p se e solo se $\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)$ è congruente a 0 o $1 \pmod q$. Secondo la Proposizione 1, questo avviene solo se $\frac{1}{4}(\langle \boldsymbol{x}, \boldsymbol{y}\rangle + 2)$ vale 0 o 1, ovvero se $\langle \boldsymbol{x}, \boldsymbol{y}\rangle \in \{-2, +2\}$. Dunque $\boldsymbol{x}$ e $\boldsymbol{y}$ devono essere quasi ortogonali, il che contraddice la definizione di Q'.

Proposizione 3. *Lo stesso risultato vale per i polinomi $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ nelle $n-1$ variabili $x_2,\ldots,x_n$ che si ottengono come segue: si scomponga $F_{\boldsymbol{y}}(\boldsymbol{x})$ in monomi; si rimuova la variabile x_1 e si riducano tutte le potenze più elevate delle altre variabili sostituendo $x_1 = 1$ e $x_i^2 = 1$ con $i > 1$. I polinomi $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ hanno grado massimo $q-2$.*

I vettori $\boldsymbol{x} \in Q \subseteq \{+1,-1\}^n$ soddisfano tutti $x_1 = 1$ e $x_i^2 = 1$. Pertanto le sostituzioni non cambiano i valori dei polinomi nell'insieme Q. Inoltre esse non aumentano il grado, quindi $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ ha grado massimo $q-2$.

Proposizione 4. *Non vi è alcuna relazione lineare (con coefficienti razionali) tra i polinomi $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$, ovvero i polinomi $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$, $\boldsymbol{y} \in Q'$ sono linearmente indipendenti su $\mathbb{Q}$. In particolare, essi sono distinti.*

Si supponga che vi sia una relazione della forma $\sum_{\boldsymbol{y}\in Q'} \alpha_{\boldsymbol{y}} \overline{F}_{\boldsymbol{y}}(\boldsymbol{x}) = 0$ tale che non tutti i coefficienti $\alpha_{\boldsymbol{y}}$ valgano zero. Dopo moltiplicazione con uno scalare appropriato possiamo supporre che tutti i coefficienti siano interi, ma non tutti siano divisibili per p. Ma allora per ogni $\boldsymbol{y} \in Q'$ la stima $\boldsymbol{x} := \boldsymbol{y}$ dà che $\alpha_{\boldsymbol{y}} \overline{F}_{\boldsymbol{y}}(\boldsymbol{y})$ è divisibile per p e pertanto lo è anche $\alpha_{\boldsymbol{y}}$, poiché $\overline{F}_{\boldsymbol{y}}(\boldsymbol{y})$ non lo è.

Proposizione 5. *$|Q'|$ è limitato dal numero di monomi senza quadrati di grado massimo $q-2$ in $n-1$ variabili, ovvero da $\sum_{i=0}^{q-2} \binom{n-1}{i}$.*

Per costruzione i polinomi $\overline{F}_{\boldsymbol{y}}$ sono privi di quadrati: nessuno dei monomi contiene una variabile di grado più alto di 1. Pertanto ogni $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ è una combinazione lineare di monomi senza quadrati di grado massimo $q-2$ nelle $n-1$ variabili $x_2,\ldots,x_n$. Dal momento che i polinomi $\overline{F}_{\boldsymbol{y}}(\boldsymbol{x})$ sono linearmente indipendenti, il loro numero (che è $|Q'|$) non può essere più grande del numero di monomi in questione.

Dettagli di (2): La prima colonna di $\boldsymbol{x}\boldsymbol{x}^T$ è $\boldsymbol{x}$. Pertanto per $\boldsymbol{x} \in Q$ distinti, otteniamo matrici distinte $M(\boldsymbol{x}) := \boldsymbol{x}\boldsymbol{x}^T$. Interpretiamo queste matrici come vettori di lunghezza n^2 con componenti $x_i x_j$. Un semplice calcolo

$$\begin{aligned}\big\langle M(\boldsymbol{x}), M(\boldsymbol{y})\big\rangle &= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i x_j)(y_i y_j) \\ &= \Big(\sum_{i=1}^{n} x_i y_i\Big)\Big(\sum_{j=1}^{n} x_j y_j\Big) \;=\; \langle \boldsymbol{x}, \boldsymbol{y}\rangle^2 \;\geq\; 4\end{aligned}$$

mostra che il prodotto scalare di $M(\boldsymbol{x})$ ed $M(\boldsymbol{y})$ è minimizzato se e solo se $\boldsymbol{x}, \boldsymbol{y} \in Q$ sono quasi ortogonali.

Dettagli di (3): Sia $U(\boldsymbol{x}) \in \{+1,-1\}^d$ il vettore di tutti gli elementi sotto la diagonale di $M(\boldsymbol{x})$. Dal momento che $M(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^T$ è simmetrico

$$M(\boldsymbol{x}) \;=\; \begin{pmatrix} 1 & & U(\boldsymbol{x}) \\ & 1 & \\ & \ddots & \\ U(\boldsymbol{x}) & & 1 \end{pmatrix}$$

con valori diagonali di $+1$, vediamo che $M(\boldsymbol{x}) \neq M(\boldsymbol{y})$ implica $U(\boldsymbol{x}) \neq U(\boldsymbol{y})$. Inoltre

$$4 \;\leq\; \langle M(\boldsymbol{x}), M(\boldsymbol{y})\rangle \;=\; 2\langle U(\boldsymbol{x}), U(\boldsymbol{y})\rangle + n,$$

ovvero

$$\langle U(\boldsymbol{x}), U(\boldsymbol{y})\rangle \;\geq\; -\frac{n}{2} + 2,$$

con uguaglianza se e solo se $\boldsymbol{x}$ e $\boldsymbol{y}$ sono quasi ortogonali. Poiché tutti i vettori $U(\boldsymbol{x}) \in S$ hanno la stessa lunghezza $\sqrt{\langle U(\boldsymbol{x}), U(\boldsymbol{x})\rangle} = \sqrt{\binom{n}{2}}$, ciò significa che la distanza massima tra i punti $U(\boldsymbol{x}), U(\boldsymbol{y}) \in S$ si ottiene esattamente quando $\boldsymbol{x}$ e $\boldsymbol{y}$ sono quasi ortogonali.

Dettagli di (4): Per $q = 9$ abbiamo $g(9) \approx 758.31$, che è maggiore di $d+1 = \binom{34}{2} + 1 = 562$.
Per ottenere un limite generale per d grandi usiamo la monotonicità e l'unimodalità dei coefficienti binomiali e le stime $n! > e(\frac{n}{e})^n$ e $n! < en(\frac{n}{e})^n$ (si veda all'appendice del Capitolo 2) e deriviamo

$$\sum_{i=0}^{q-2}\binom{4q-3}{i} < q\binom{4q}{q} = q\,\frac{(4q)!}{q!(3q)!} < q\,\frac{e\,4q\left(\frac{4q}{e}\right)^{4q}}{e\left(\frac{q}{e}\right)^{q} e\left(\frac{3q}{e}\right)^{3q}} = \frac{4q^2}{e}\Big(\frac{256}{27}\Big)^q.$$

Concludiamo pertanto

$$f(d) \;\geq\; g(q) \;=\; \frac{2^{4q-4}}{\sum_{i=0}^{q-2}\binom{4q-3}{i}} \;>\; \frac{e}{64q^2}\Big(\frac{27}{16}\Big)^q.$$

Da questo, con

$$d \;=\; (2q-1)(4q-3) \;=\; 5q^2 + (q-3)(3q-1) \;\geq\; 5q^2 \quad \text{per } q \geq 3,$$

$$q = \tfrac{5}{8} + \sqrt{\tfrac{d}{8} + \tfrac{1}{64}} > \sqrt{\tfrac{d}{8}}, \qquad \text{e} \qquad \left(\tfrac{27}{16}\right)^{\sqrt{\frac{1}{8}}} > 1.2032,$$

otteniamo

$$f(d) \;>\; \frac{e}{13d}(1.2032)^{\sqrt{d}} \;>\; (1.2)^{\sqrt{d}} \quad \text{per ogni } d \text{ grande abbastanza.}\ \square$$

Un controesempio di dimensione 560 si ottiene notando che per $q = 9$ il quoziente $g(q) \approx 758$ è *molto* maggiore della dimensione $d(q) = 561$. Pertanto un controesempio per $d = 560$ si ottiene prendendo solo i tre quarti dei punti in S che soddisfano $x_{21} + x_{31} + x_{32} = -1$.
La congettura di Borsuk è nota per essere vera per $d \leq 3$, ma non è stata dimostrata per nessuna dimensione più grande. Tuttavia, essa *è* vera fino a $d = 8$ se ci limitiamo ai sottoinsiemi $S \subseteq \{1, -1\}^d$, come nella costruzione precedente (si veda [8]). In entrambi i casi è possibile che si trovino controesempi in dimensioni ragionevolmente piccole.

Bibliografia

[1] K. BORSUK: *Drei Sätze über die n-dimensionale euklidische Sphäre,* Fundamenta Math. **20** (1933), 177-190.

[2] A. HINRICHS & C. RICHTER: *New sets with large Borsuk numbers,* Discrete Math., **270** (2003), 136-146.

[3] J. KAHN & G. KALAI: *A counterexample to Borsuk's conjecture,* Bulletin Amer. Math. Soc. **29** (1993), 60-62.

[4] A. NILLI: *On Borsuk's problem,* in: "Jerusalem Combinatorics '93" (H. Barcelo and G. Kalai, eds.), Contemporary Mathematics **178**, Amer. Math. Soc. 1994, 209-210.

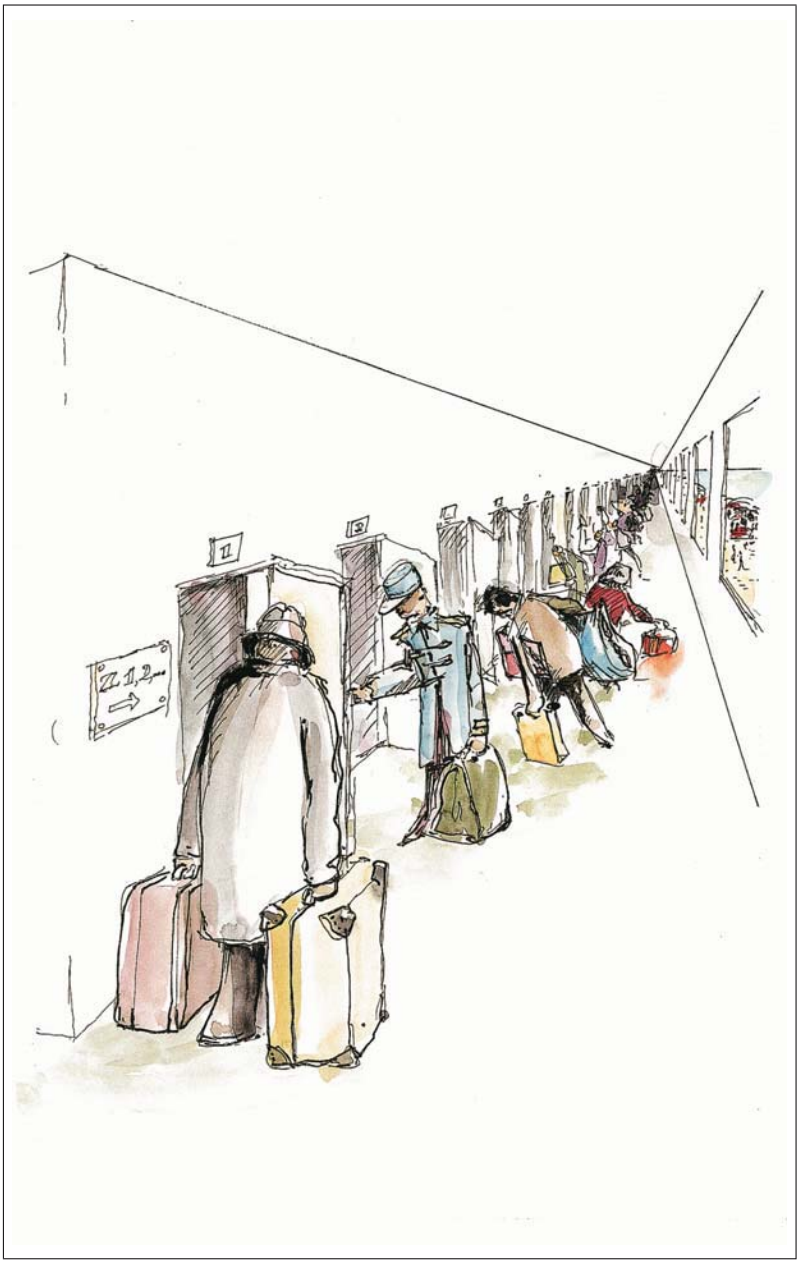
[5] A. M. RAIGORODSKII: *On the dimension in Borsuk's problem,* Russian Math. Surveys (6) **52** (1997), 1324-1325.

[6] O. SCHRAMM: *Illuminating sets of constant width,* Mathematika **35** (1988), 180-199.

[7] B. WEISSBACH: *Sets with large Borsuk number,* Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry **41** (2000), 417-423.

[8] G. M. ZIEGLER: *Coloring Hamming graphs, optimal binary codes, and the 0/1-Borsuk problem in low dimensions,* Lecture Notes in Computer Science **2122**, Springer-Verlag 2001, 164-175.

Analisi

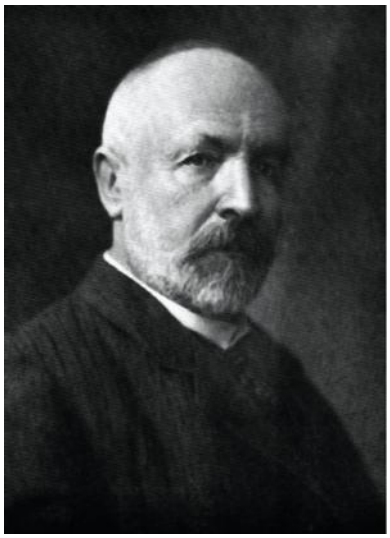


"L'hotel in riva al mare di Hilbert"

Insiemi, funzioni e l'ipotesi del continuo

Capitolo 16

La teoria degli insiemi, fondata da Georg Cantor nella seconda metà del diciannovesimo secolo, ha profondamente trasformato la matematica. La matematica dei tempi moderni è impensabile senza il concetto di insieme, ovvero, come afferma David Hilbert, "Nessuno ci scaccerà dal paradiso (della teoria degli insiemi) che Cantor ha creato per noi".



Georg Cantor

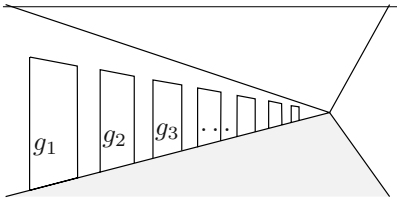
Uno dei concetti basilari proposti da Cantor fu la nozione di *dimensione* o *cardinalità* di un insieme M, indicata con $|M|$. Per gli insiemi finiti, ciò non presenta difficoltà: basta contare il numero di elementi e dire che M è un n-insieme o ha dimensione n, se M contiene esattamente n elementi. Pertanto due insiemi finiti M e N hanno dimensione uguale, $|M| = |N|$, se essi contengono lo stesso numero di elementi.

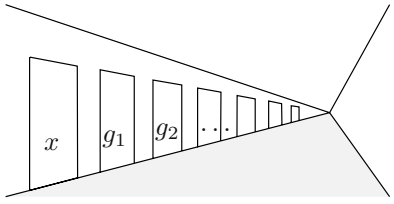
Per portare questa nozione di *uguale* dimensione agli insiemi infiniti, usiamo il seguente esperimento suggestivo per gli insiemi finiti. Supponiamo che un certo numero di persone salga su un autobus. Quando diremo che il numero di passeggeri è lo stesso del numero di sedili? Semplicemente, facciamo sedere tutti i passeggeri. Se ognuno trova un posto e nessun sedile resta vuoto, allora e soltanto allora i due insiemi (dei passeggeri e dei posti) si accordano in numero. In altre parole, le due dimensioni sono le stesse se vi è *biiezione* di un insieme sull'altro.

Questa è dunque la nostra definizione: due insiemi arbitrari M e N (finiti o infiniti) sono detti di *uguale dimensione* o *cardinalità*, se e solo se esiste una biiezione da M su N. Chiaramente, questa nozione di uguale dimensione è una relazione di equivalenza e possiamo pertanto associare un numero, detto *numero cardinale*, ad ogni classe di insiemi di uguale dimensione. Ad esempio, otteniamo per insiemi finiti i numeri cardinali $0, 1, 2, \ldots, n, \ldots$ dove n indica la classe di n-insiemi e in particolare, 0 indica l'*insieme vuoto* $\varnothing$. Osserviamo inoltre il fatto evidente che un sottoinsieme proprio di un insieme finito M ha invariabilmente dimensione inferiore a quella di M.

La teoria diventa molto interessante (e tutt'altro che intuitiva) quando ci interessiamo agli insiemi infiniti. Si consideri l'insieme $\mathbb{N} = \{1, 2, 3, \ldots\}$ dei numeri naturali. Definiamo un insieme M *numerabile* se esso può essere messo in corrispondenza biunivoca con $\mathbb{N}$. In altre parole, M è numerabile se possiamo elencare gli elementi di M come $m_1, m_2, m_3, \ldots$. Ma a questo punto accade un fenomeno strano. Supponiamo di aggiungere ad $\mathbb{N}$ un nuovo elemento x. Allora $\mathbb{N} \cup \{x\}$ è ancora numerabile e pertanto ha la stessa dimensione di $\mathbb{N}$!

Questo fatto è deliziosamente illustrato dall'"hotel di Hilbert". Supponia-



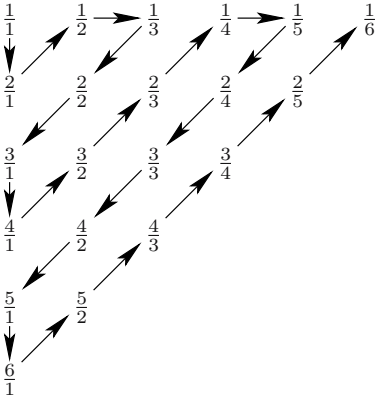


mo che un hotel abbia un'infinità numerabile di stanze, numerate $1, 2, 3, \ldots$ tali che il cliente g_i occupi la stanza i; l'hotel è completo. Ora arriva un nuovo cliente x chiedendo una stanza, al che il direttore dell'albergo risponde: spiacente, tutte le stanze sono occupate. Nessun problema, replica il nuovo arrivato, sposti il cliente g_1 nella stanza 2, g_2 nella 3, g_3 nella 4, e così via, ed io prenderò la stanza 1. Con sorpresa del direttore (il quale non è un matematico) funziona: può ancora sistemare tutti i clienti più il nuovo arrivato x!

Ora è chiaro che egli può anche sistemare un altro cliente y ed un altro ancora, z, e così via. In particolare, notiamo che, contrariamente al caso degli insiemi finiti, può ben darsi che un sottoinsieme proprio di un insieme *infinito* M abbia la stessa dimensione di M. In effetti, come vedremo, questa è una definizione dell'infinito: un insieme è infinito se e solo se esso ha la stessa dimensione di uno dei suoi sottoinsiemi propri.

Lasciamo adesso l'hotel di Hilbert e guardiamo ai nostri familiari insiemi di numeri. L'insieme $\mathbb{Z}$ dei numeri interi è ancora numerabile, poiché possiamo enumerare $\mathbb{Z}$ nella forma $\mathbb{Z} = \{0, 1, -1, 2, -2, 3, -3, \ldots\}$. Può sembrare più sorprendente che i numeri razionali possano essere enumerati nello stesso modo.

Teorema 1. *L'insieme $\mathbb{Q}$ dei numeri razionali è numerabile.*



■ **Dimostrazione.** Elencando gli elementi dell'insieme dei numeri razionali positivi $\mathbb{Q}^+$ come suggerito nella figura a margine, ma tralasciando i numeri già trovati, vediamo che $\mathbb{Q}^+$ è numerabile e pertanto lo è anche $\mathbb{Q}$ elencando 0 all'inizio e $-\frac{p}{q}$ appena dopo $\frac{p}{q}$. Con questa elencazione

$$\mathbb{Q} = \{0, 1, -1, 2, -2, \tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{3}, -\tfrac{1}{3}, 3, -3, 4, -4, \tfrac{3}{2}, -\tfrac{3}{2}, \ldots\}. \qquad \square$$

Un altro modo di interpretare la figura è fornito dalla seguente proposizione:

> *L'unione di un insieme numerabile di insiemi numerabili M_n è a sua volta numerabile.*

In effetti, si ponga $M_n = \{a_{n1}, a_{n2}, a_{n3}, \ldots\}$ e si elenchi

$$\bigcup_{n=1}^{\infty} M_n = \{a_{11}, a_{21}, a_{12}, a_{13}, a_{22}, a_{31}, a_{41}, a_{32}, a_{23}, a_{14}, \ldots\}$$

esattamente come prima.

Analizziamo ora un po' meglio l'enumerazione di Cantor dei numeri razionali positivi. Guardando la figura abbiamo ottenuto la successione

$$\tfrac{1}{1}, \tfrac{2}{1}, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{2}{2}, \tfrac{3}{1}, \tfrac{4}{1}, \tfrac{3}{2}, \tfrac{2}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \tfrac{2}{4}, \tfrac{3}{3}, \tfrac{4}{2}, \tfrac{5}{1}, \ldots$$

e dunque abbiamo dovuto eliminare i doppioni come $\frac{2}{2} = \frac{1}{1}$ oppure $\frac{2}{4} = \frac{1}{2}$.

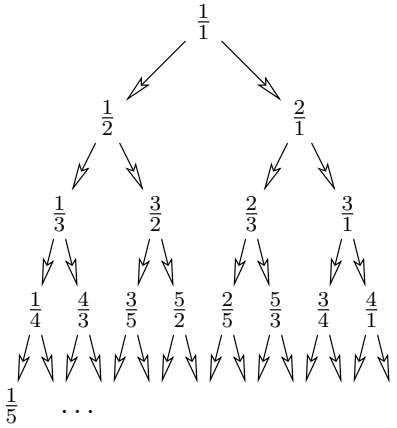
Ma vi è un'elencazione ancora più elegante e sistematica, priva di doppioni, trovata abbastanza recentemente da Neil Calkin e Herbert Wilf. Il nuovo elenco inizia come segue:

$$\tfrac{1}{1},\ \tfrac{1}{2},\ \tfrac{2}{1},\ \tfrac{1}{3},\ \tfrac{3}{2},\ \tfrac{2}{3},\ \tfrac{3}{1},\ \tfrac{1}{4},\ \tfrac{4}{3},\ \tfrac{3}{5},\ \tfrac{5}{2},\ \tfrac{2}{5},\ \tfrac{5}{3},\ \tfrac{3}{4},\ \tfrac{4}{1},\ \ldots\ .$$

Qui il denominatore dell'n-esimo numero razionale è uguale al numeratore dell'$(n+1)$-esimo. In altre parole, la frazione n-esima è $b(n)/b(n+1)$, dove $\big(b(n)\big)_{n\geq 0}$ è una successione che inizia con

$$(1,\ 1,\ 2,\ 1,\ 3,\ 2,\ 3,\ 1,\ 4,\ 3,\ 5,\ 2,\ 5,\ 3,\ 4,\ 1,\ 5,\ \ldots).$$

Questa successione è stata studiata per la prima volta da un matematico tedesco, Moritz Abraham Stern, in una pubblicazione del 1858, ed è divenuta nota come "la serie diatomica di Stern". Come si ottiene questa successione e dunque l'elencazione di Calkin-Wilf delle frazioni positive? Si consideri l'albero binario infinito a margine. Notiamo immediatamente la sua legge ricorsiva:



- $\frac{1}{1}$ è al vertice dell'albero,
- ogni nodo $\frac{i}{j}$ ha due figli: il figlio di sinistra è $\frac{i}{i+j}$ e quello di destra è $\frac{i+j}{j}$.

Possiamo verificare con facilità le quattro proprietà seguenti:

(1) Tutte le frazioni nell'albero sono ridotte, ovvero, se $\frac{r}{s}$ compare nell'albero, allora r ed s sono primi fra loro.

Questo vale per il vertice $\frac{1}{1}$ e quindi procediamo per induzione verso il basso. Se r ed s sono primi fra loro, allora lo sono anche r ed $r+s$, e così anche s ed $r+s$.

(2) Ogni frazione ridotta $\frac{r}{s} > 0$ è presente nell'albero.

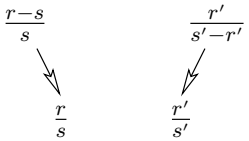
Usiamo il procedimento di induzione sulla somma $r+s$. Il più piccolo valore è $r+s=2$, ovvero $\frac{r}{s}=\frac{1}{1}$, e questo appare al vertice. Se $r>s$, allora per induzione $\frac{r-s}{s}$ compare nell'albero, e pertanto otteniamo $\frac{r}{s}$ come figlio di destra. Analogamente se $r<s$, allora appare $\frac{r}{s-r}$, che ha come figlio di sinistra $\frac{r}{s}$.

(3) Ogni frazione ridotta compare esattamente una volta.

L'argomentazione è analoga. Se $\frac{r}{s}$ compare più di una volta, allora $r \neq s$, poiché ogni nodo nell'albero, tranne il vertice, è della forma $\frac{i}{i+j} < 1$ ovvero $\frac{i+j}{j} > 1$. Ma se $r>s$ o $r<s$, allora usiamo lo stesso argomento induttivo come prima.

Ogni numero razionale positivo compare pertanto esattamente una volta nell'albero, e possiamo scrivere tali numeri elencandoli livello per livello da sinistra a destra. Questo dà precisamente il segmento iniziale mostrato sopra.

(4) Il denominatore dell'n-esima frazione nella nostra lista è uguale al numeratore dell'$(n+1)$-esima.



Ciò è certamente vero per $n = 0$, ovvero quando la frazione n-esima è un figlio sinistro. Supponiamo che l'n-esimo numero $\frac{r}{s}$ sia un figlio destro. Se $\frac{r}{s}$ è alla frontiera destra, allora $s = 1$, ed il successore giace alla frontiera sinistra ed ha numeratore 1. Infine, se $\frac{r}{s}$ è all'interno e $\frac{r'}{s'}$ è la frazione seguente nella nostra successione, allora $\frac{r}{s}$ è il figlio destro di $\frac{r-s}{s}$, $\frac{r'}{s'}$ è il figlio sinistro di $\frac{r'}{s'-r'}$ e per induzione il denominatore di $\frac{r-s}{s}$ è il numeratore di $\frac{r'}{s'-r'}$, dunque otteniamo $s = r'$.

Questo è certamente interessante, ma c'è ancora di più. Vi sono due domande naturali:

– La successione $\big(b(n)\big)_{n\geq 0}$ ha un "senso"? Ovvero, $b(n)$ conta qualcosa di semplice?

– Dato $\frac{r}{s}$, esiste un modo semplice per determinare il successivo nell'elenco?

Per rispondere alla prima domanda, troviamo che il nodo $b(n)/b(n+1)$ ha come figli $b(2n+1)/b(2n+2)$ e $b(2n+2)/b(2n+3)$. In base alla configurazione dell'albero otteniamo le relazioni ricorsive

$$b(2n+1) = b(n) \quad \text{e} \quad b(2n+2) = b(n) + b(n+1). \tag{1}$$

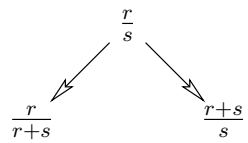
Posto $b(0) = 1$ la successione $(b(n))_{n\geq 0}$ è completamente determinata da (1).

Esiste dunque una successione nota interessante che obbedisca alla stessa relazione ricorsiva? Sì, esiste. Sappiamo che ogni numero n può essere scritto in modo univoco come somma di potenze distinte di 2 — questa è la solita rappresentazione binaria di n. Una rappresentazione *iper-binaria* di n è una rappresentazione di n come somma di potenze di 2, dove ogni potenza 2^k appare al massimo *due volte*. Sia $h(n)$ il numero di tali rappresentazioni di n. Siete invitati a verificare che la successione $h(n)$ soddisfa (1) e questo dà $b(n) = h(n)$ per ogni n.

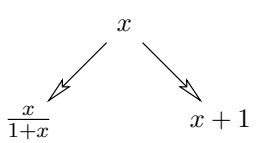
Ad esempio, $h(6) = 3$, con le rappresentazioni iperbinarie
$6 = 4 + 2$
$6 = 4 + 1 + 1$
$6 = 2 + 2 + 1 + 1$.

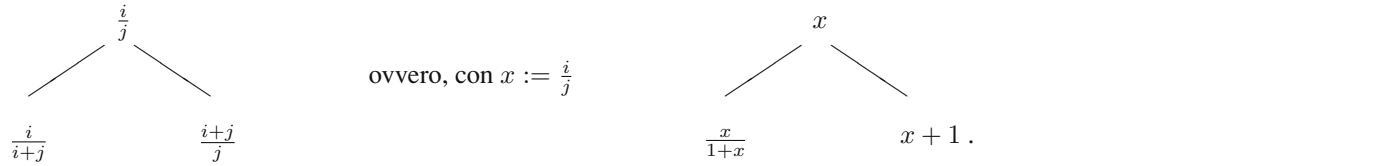
Incidentalmente, abbiamo dimostrato un fatto sorprendente: se $\frac{r}{s}$ è una frazione ridotta, allora esiste esattamente un numero intero n con $r = h(n)$ ed $s = h(n+1)$.

Diamo uno sguardo alla seconda domanda. Nel nostro albero abbiamo

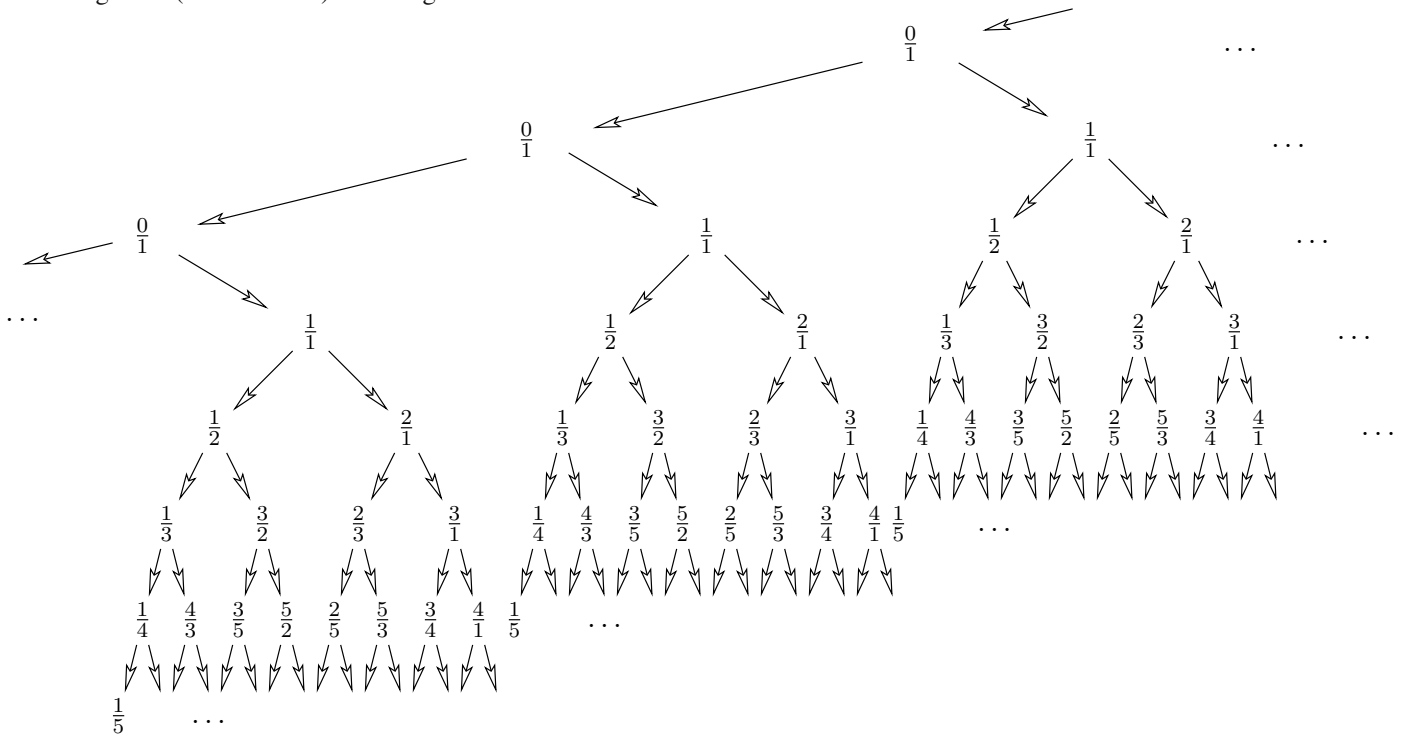


ovvero, con $x := \frac{r}{s}$,

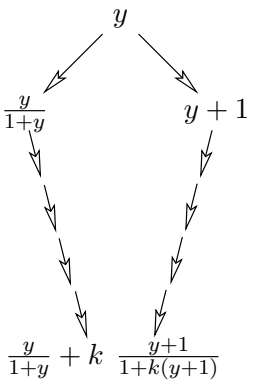




Su questa base possiamo generare un albero binario infinito ancora più grande (senza vertice) come segue:



In questo albero tutte le righe sono uguali, e tutte mostrano l'elencazione di Calkin-Wilf dei numeri razionali positivi (iniziando con uno $\frac{0}{1}$ addizionale). Allora come si passa da un numero razionale al successivo? Per rispondere a questa domanda, notiamo dapprima che il figlio di destra di ogni numero razionale x è $x+1$, il nipote di destra è $x+2$, pertanto il figlio di destra dopo la k-esima generazione è $x+k$. Analogamente, il figlio di sinistra di x è $\frac{x}{1+x}$, il cui figlio di sinistra è $\frac{x}{1+2x}$, e così via: il discendente di sinistra di x preso dopo k generazioni è $\frac{x}{1+kx}$. Ora per trovare il modo di passare da $\frac{r}{s} = x$ al numero razionale "seguente" $f(x)$ nell'elenco dobbiamo analizzare la situazione rappresentata a margine. In effetti, se consideriamo ogni numero razionale non negativo x nel nostro albero binario infinito, allora esso è il discendente di destra dopo k generazioni del figlio di sinistra di un numero razionale $y \geq 0$ (per un $k \geq 0$), mentre $f(x)$ è dato come il discendente di sinistra dopo k generazioni del figlio di destra dello stesso y. Pertanto mediante le formule per i figli di sinistra e di destra alla k-esima



generazione, otteniamo

$$x \;=\; \frac{y}{1+y} + k,$$

come illustrato nella figura a margine alla pagina precedente. Qui $k = \lfloor x \rfloor$ è la parte intera di x, mentre $\frac{y}{1+y} = \{x\}$ è la sua parte frazionaria. Da ciò otteniamo

$$f(x) = \frac{y+1}{1+k(y+1)} = \frac{1}{\frac{1}{y+1}+k} = \frac{1}{k+1-\frac{y}{y+1}} = \frac{1}{\lfloor x \rfloor + 1 - \{x\}}\,.$$

Siamo pertanto giunti ad una splendida formula per il successore $f(x)$ di x, trovata molto recentemente da Moshe Newman:

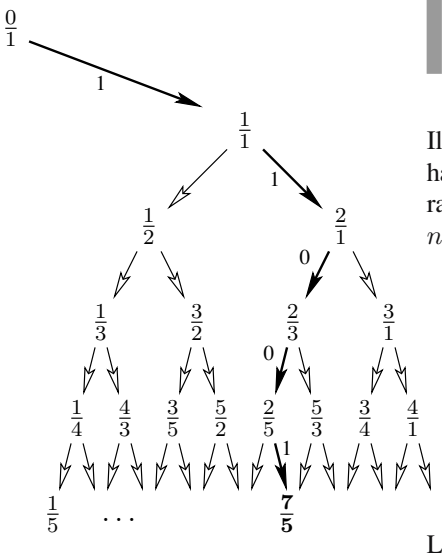
La funzione

$$x \;\longmapsto\; f(x) \;=\; \frac{1}{\lfloor x \rfloor + 1 - \{x\}}$$

genera la successione di Calkin-Wilf

$$\tfrac{1}{1} \mapsto \tfrac{1}{2} \mapsto \tfrac{2}{1} \mapsto \tfrac{1}{3} \mapsto \tfrac{3}{2} \mapsto \tfrac{2}{3} \mapsto \tfrac{3}{1} \mapsto \tfrac{1}{4} \mapsto \tfrac{4}{3} \mapsto \ldots$$

che contiene ogni numero razionale positivo solo una volta.



Il metodo Calkin-Wilf-Newman per enumerare i numeri razionali positivi ha numerose altre proprietà notevoli. Ad esempio, si può chiedere un modo rapido per determinare la frazione ennesima nella successione, diciamo per $n = 10^6$. Eccolo:

> Per trovare la frazione ennesima nella successione di Calkin-Wilf, si esprima n come un numero binario $n = (b_k b_{k-1} ... b_1 b_0)_2$ e quindi si segua il cammino nell'albero di Calkin-Wilf determinato dalle sue cifre, iniziando da $\frac{s}{t} = \frac{0}{1}$.
> Qui $b_i = 1$ significa "prendere il figlio di destra", ovvero "sommare il denominatore ed il numeratore", mentre $b_i = 0$ significa "prendere il figlio di sinistra", ovvero, "sommare il numeratore ed il denominatore".

La figura a margine mostra il cammino risultante per $n = 25 = (11001)_2$: pertanto il 25-esimo numero nella successione di Calkin-Wilf è $\frac{7}{5}$. Il lettore potrà trovare uno schema simile che calcola per una frazione data $\frac{s}{t}$ la (rappresentazione binaria della) sua posizione n nella successione di Calkin-Wilf.

Passiamo ora ai numeri reali $\mathbb{R}$. Sono anch'essi numerabili? No, non lo sono ed il metodo con cui lo si dimostra — il *metodo di diagonalizzazione* di Cantor — non solo è di fondamentale importanza per tutta la teoria degli insiemi, ma appartiene certamente al *Libro* dal momento che si tratta di un raro colpo di genio.

Teorema 2. *L'insieme $\mathbb{R}$ dei numeri reali **non** è numerabile.*

■ **Dimostrazione.** Ogni sottoinsieme N di un insieme numerabile $M = \{m_1, m_2, m_3, \ldots\}$ è *al più numerabile* (ovvero, finito o numerabile). In effetti, basta elencare gli elementi di N come appaiono in M. Pertanto, se possiamo trovare un sottoinsieme di $\mathbb{R}$ che non sia numerabile, allora a fortiori $\mathbb{R}$ non può essere numerabile. L'insieme M di $\mathbb{R}$ che vogliamo analizzare è l'intervallo $(0, 1]$ di tutti i numeri reali positivi r con $0 < r \leq 1$. Procediamo per contraddizione supponendo che M sia numerabile e sia $M = \{r_1, r_2, r_3, \ldots\}$ un'elencazione di M. Scriviamo r_n attraverso il suo unico sviluppo decimale *infinito* senza una successione infinita di zeri alla fine:

$$r_n = 0.a_{n1}a_{n2}a_{n3}...$$

dove $a_{ni} \in \{0, 1, \ldots, 9\}$ per ogni n ed i. Ad esempio, $0.7 = 0.6999...$ Si consideri ora la matrice doppiamente infinita

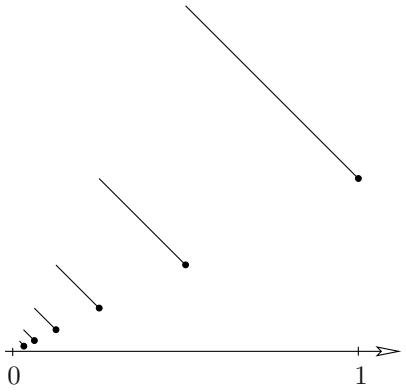
$$\begin{array}{rcl} r_1 & = & 0.a_{11}a_{12}a_{13}... \\ r_2 & = & 0.a_{21}a_{22}a_{23}... \\ \vdots & & \vdots \\ r_n & = & 0.a_{n1}a_{n2}a_{n3}... \\ \vdots & & \vdots \end{array}$$

Per ogni n, si scelga $b_n \in \{1, \ldots, 8\}$ diverso da a_{nn}; chiaramente questo è fattibile. Allora $b = 0.b_1b_2b_3...b_n...$ è un numero reale nel nostro insieme M e pertanto deve avere un indice, diciamo $b = r_k$. Ma ciò non può essere, dal momento che b_k è diverso da a_{kk}. E questa è la dimostrazione completa! □

Soffermiamoci sui numeri reali per un momento. Notiamo che tutti e quattro i tipi di intervallo $(0, 1), (0, 1], [0, 1)$ e $[0, 1]$ hanno la stessa dimensione. Come esempio, verifichiamo che $(0, 1]$ e $(0, 1)$ hanno uguale cardinalità. La trasformazione $f : (0, 1] \longrightarrow (0, 1), x \longmapsto y$ definita da

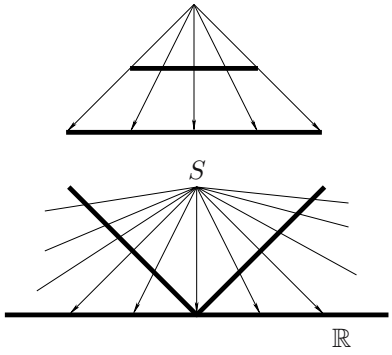
$$y := \begin{cases} \frac{3}{2} - x & \text{per} \quad \frac{1}{2} < x \leq 1, \\ \frac{3}{4} - x & \text{per} \quad \frac{1}{4} < x \leq \frac{1}{2}, \\ \frac{3}{8} - x & \text{per} \quad \frac{1}{8} < x \leq \frac{1}{4}, \\ \quad \vdots & \end{cases}$$

fa al caso nostro. In effetti, essa è biiettiva, dal momento che l'intervallo descritto da y nella prima riga è $\frac{1}{2} \leq y < 1$, nella seconda riga $\frac{1}{4} \leq y < \frac{1}{2}$, nella terza riga $\frac{1}{8} \leq y < \frac{1}{4}$, e così via.



Una $f : (0, 1] \longrightarrow (0, 1)$ biiettiva

In seguito troviamo che *ogni* coppia di intervalli (di lunghezza finita > 0) ha dimensione uguale considerando la proiezione centrale come nella figura. Ed è vero anche di più: ogni intervallo (di lunghezza > 0) ha la stessa dimensione dell'intera retta $\mathbb{R}$. Per verificarlo, si guardi all'intervallo curvo aperto $(0, 1)$ e si proietti su $\mathbb{R}$ a partire dal centro S.



Pertanto, in conclusione, ogni intervallo aperto, semi-aperto o chiuso (finito o infinito) di lunghezza > 0 ha la stessa dimensione, che denotiamo con c, dove c sta per *continuo* (un termine talvolta usato per l'intervallo [0,1]).

Il fatto che tali intervalli finiti ed infiniti abbiano la stessa dimensione può risultare prevedibile ripensandoci, ma è un fatto chiaramente non intuitivo.

Teorema 3. *L'insieme $\mathbb{R}^2$ di tutte la coppie ordinate di numeri reali (ovvero, il piano reale) ha la stessa dimensione di $\mathbb{R}$.*

■ **Dimostrazione.** Per verificarlo basta dimostrare che l'insieme di tutte le coppie (x, y), $0 < x, y \leq 1$, può essere trasformato biiettivamente su $(0, 1]$. La dimostrazione proviene ancora dal *Libro*. Si consideri la coppia (x, y) e si scrivano x e y sotto forma dei loro sviluppi decimali illimitati come nel seguente esempio:

$$\begin{array}{rcccccccc} x & = & 0.3 & & 01 & 2 & 007 & 08 & \ldots \\ y & = & 0.009 & 2 & 05 & 1 & & 0008 & \ldots \end{array}$$

Si noti che abbiamo separato le cifre di x ed y in gruppi che includono alla loro fine la cifra non nulla successiva. Ora associamo a (x, y) il numero $z \in (0, 1]$ scrivendo il primo x-gruppo, quindi il primo y-gruppo, in seguito il secondo x-gruppo e così via. Dunque, nel nostro esempio, otteniamo

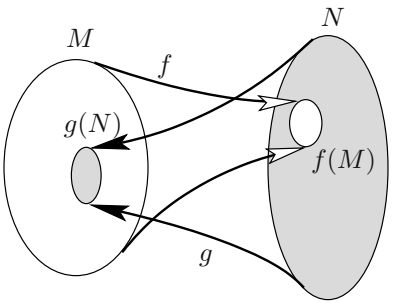
$$z \;=\; 0.3\,009\,01\,2\,2\,05\,007\,1\,08\,0008\;\ldots$$

Poiché né x né y presentano solo zeri a partire da un certo punto, troviamo che l'espressione per z è nuovamente uno sviluppo decimale illimitato. In modo speculare, dallo sviluppo di z possiamo immediatamente ricostruire la preimmagine (x, y), dunque la trasformazione è biiettiva — fine della dimostrazione. □

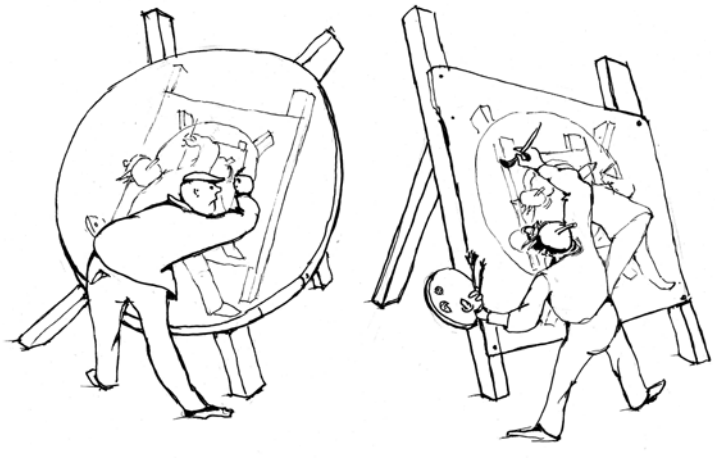
Dal momento che $(x, y) \longmapsto x+iy$ è una biiezione da $\mathbb{R}^2$ sui numeri complessi $\mathbb{C}$, concludiamo che $|\mathbb{C}| = |\mathbb{R}| = c$. Perché il risultato $|\mathbb{R}^2| = |\mathbb{R}|$ è tanto imprevedibile? Perché va contro la nostra intuizione di *dimensione*. Esso afferma che il piano bidimensionale $\mathbb{R}^2$ (e in generale, per induzione, lo spazio n-dimensionale $\mathbb{R}^n$) può essere trasformato biiettivamente sulla retta monodimensionale $\mathbb{R}$. Pertanto la dimensione non è generalmente conservata dalle trasformazioni biiettive. Se, tuttavia, esigiamo che la trasformazione e la sua inversa siano continue, allora la dimensione è conservata, come fu dimostrato per la prima volta da Luitzen Brouwer.

Andiamo ora un po' più lontano. Finora, abbiamo acquisito la nozione di uguale cardinalità. Quando potremo dire che M è grande al massimo quanto N? Le trasformazioni forniscono nuovamente la chiave. Diciamo che il numero cardinale $\mathfrak{m}$ è *inferiore o uguale a* $\mathfrak{n}$, se per due insiemi M e N con $|M| = \mathfrak{m}$, $|N| = \mathfrak{n}$, esiste una *iniezione* da M in N. Chiaramente, la relazione $\mathfrak{m} \leq \mathfrak{n}$ è indipendente dagli insiemi rappresentativi scelti, M e N. Per gli insiemi finiti ciò corrisponde di nuovo alla nostra nozione intuitiva: un m-insieme è al più grande quanto un n-insieme se e solo se $m \leq n$.

Ora dobbiamo affrontare un problema di base. Certamente ci piacerebbe che le leggi consuete sulle disuguaglianze valessero anche per i numeri cardinali. Ma questo è vero quando si tratta di infiniti numeri cardinali? In particolare, è vero che $\mathfrak{m} \leq \mathfrak{n}$, $\mathfrak{n} \leq \mathfrak{m}$ implica $\mathfrak{m} = \mathfrak{n}$? Ciò non è affatto ovvio: supponiamo di avere infiniti insiemi M e N ed anche trasformazioni $f : M \longrightarrow N$ e $g : N \longrightarrow M$ che siano iniettive ma non necessariamente suriettive. Questo suggerisce di costruire una biiezione mettendo in relazione alcuni elementi $m \in M$ con $f(m) \in N$ ed alcuni elementi $n \in N$ con $g(n) \in M$. Tuttavia non è chiaro se le diverse scelte possibili si possano "mettere d'accordo".



La risposta affermativa è fornita dal celebre teorema di Schröder-Bernstein, che Cantor enunciò nel 1883. Le prime dimostrazioni furono date da Friedrich Schröder e Felix Bernstein ben più tardi. La seguente dimostrazione appare in un piccolo libro di uno dei giganti della teoria degli insiemi del ventesimo secolo, Paul Cohen, celebre per aver risolto l'ipotesi del continuo (che discuteremo fra poco).

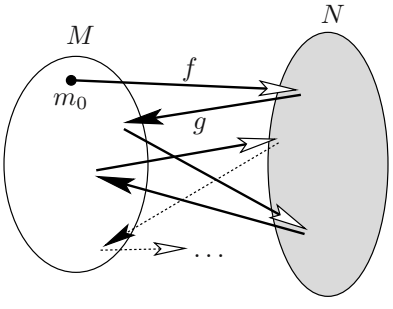


"Schröder e Bernstein che dipingono"

Teorema 4. *Se ognuno dei due insiemi M e N si può trasformare iniettivamente nell'altro, allora vi è una biiezione da M a N, ovvero, $|M| = |N|$.*

■ **Dimostrazione.** Possiamo certamente supporre che M e N siano disgiunti; in caso contrario, sostituiamo semplicemente N con una nuova copia.

Ora f e g trasformano "avanti e indietro" gli elementi di M e quelli di N. Per rendere questa situazione perfettamente chiara ed ordinata allineamo $M \cup N$ in catene di elementi: si prenda un elemento arbitrario $m_0 \in M$ e si generi da questo una catena di elementi applicando f, quindi g, poi ancora f, quindi g, e così via. La catena si chiude (si tratta del Caso 1) se raggiungiamo ancora m_0 in questo procedimento oppure può continuare indefinitamente con elementi distinti. Il primo "duplicato" nella catena non può essere un elemento diverso da m_0, per l'iniettività.



Se la catena continua indefinitamente, allora proviamo a seguirla all'indietro: da m_0 a $g^{-1}(m_0)$ se m_0 è nell'immagine di g, quindi a $f^{-1}(g^{-1}(m_0))$ se $g^{-1}(m_0)$ è nell'immagine di f e così via. Qui possono sorgere tre altri casi: il processo di seguire la catena all'indietro può proseguire indefinitamente (Caso 2), può arrestarsi su un elemento di M che non sta nell'immagine di g (Caso 3) oppure può arrestarsi ad un elemento di N che non appartiene all'immagine di f (Caso 4).

Pertanto $M \cup N$ si divide perfettamente in quattro tipi di catene, i cui elementi possiamo contrassegnare in modo tale che una biiezione sia data semplicemente ponendo $F : m_i \longmapsto n_i$. Verifichiamolo nei quattro casi separatamente:

Caso 1. Cicli finiti su $2k+2$ elementi distinti ($k \geq 0$)

$$m_0 \overset{f}{\twoheadrightarrow} n_0 \overset{g}{\rightarrow} m_1 \overset{f}{\twoheadrightarrow} \cdots \quad m_k \overset{f}{\twoheadrightarrow} n_k \quad (n_k \overset{g}{\rightarrow} m_0)$$

Caso 2. Catene infinite di elementi distinti nei due sensi

$$\cdots \rightarrow m_0 \overset{f}{\twoheadrightarrow} n_0 \overset{g}{\rightarrow} m_1 \overset{f}{\twoheadrightarrow} n_1 \overset{g}{\rightarrow} m_2 \overset{f}{\twoheadrightarrow} \cdots$$

Caso 3. Catene infinite di elementi distinti in un senso che iniziano con gli elementi $m_0 \in M \backslash g(N)$

$$m_0 \overset{f}{\twoheadrightarrow} n_0 \overset{g}{\rightarrow} m_1 \overset{f}{\twoheadrightarrow} n_1 \overset{g}{\rightarrow} m_2 \overset{f}{\twoheadrightarrow} \cdots$$

Caso 4. Catene infinite di elementi distinti in un senso che iniziano con gli elementi $n_0 \in N \backslash f(M)$

$$n_0 \overset{g}{\rightarrow} m_0 \overset{f}{\twoheadrightarrow} n_1 \overset{g}{\rightarrow} m_1 \overset{f}{\twoheadrightarrow} \cdots \quad \square$$

"Il più piccolo cardinale infinito"

Che ne è delle altre relazioni che governano le disuguaglianze? Come al solito, poniamo $\mathfrak{m} < \mathfrak{n}$ se $\mathfrak{m} \leq \mathfrak{n}$, ma $\mathfrak{m} \neq \mathfrak{n}$. Abbiamo appena visto che per ogni coppia di numeri cardinali $\mathfrak{m}$ e $\mathfrak{n}$ vale al massimo una delle tre possibilità

$$\mathfrak{m} < \mathfrak{n}, \ \mathfrak{m} = \mathfrak{n}, \ \mathfrak{m} > \mathfrak{n},$$

e segue dalla teoria dei numeri cardinali che, in effetti, esattamente una delle relazioni è vera. Si veda l'appendice a questo capitolo, la Proposizione 2.

Inoltre, il teorema di Schröder-Bernstein ci dice che la relazione $<$ è transitiva, ovvero, $\mathfrak{m} < \mathfrak{n}$ e $\mathfrak{n} < \mathfrak{p}$ implicano $\mathfrak{m} < \mathfrak{p}$. Pertanto le cardinalità sono disposte in ordine lineare iniziando dai numeri cardinali finiti $0, 1, 2, 3, \ldots$. Ricorrendo al solito sistema di assiomi di Zermelo-Fraenkel (in particolare, l'assioma della scelta) troviamo facilmente che ogni insieme infinito M contiene un sottoinsieme numerabile. In effetti, M contiene un elemento, diciamo m_1. L'insieme $M \setminus \{m_1\}$ non è vuoto (dal

momento che è infinito) e pertanto contiene un elemento m_2. Considerando $M \setminus \{m_1, m_2\}$ deduciamo l'esistenza di m_3 e così via. Pertanto, la dimensione di un insieme numerabile è il *più piccolo* cardinale *infinito*, usualmente indicato con $\aleph_0$ (pronunciato "aleph-con-zero").

Come corollario a $\aleph_0 \leq \mathfrak{m}$ per ogni cardinale infinito $\mathfrak{m}$, possiamo immediatamente dimostrare "l'hotel di Hilbert" per ogni numero cardinale infinito $\mathfrak{m}$, ovvero, che abbiamo $|M \cup \{x\}| = |M|$ per ogni insieme infinito M. In effetti, M contiene un sottoinsieme $N = \{m_1, m_2, m_3, \ldots\}$. Ora trasformiamo x su m_1, m_1 su m_2 e così via, mantenendo fissi gli elementi di $M \setminus N$. Da qui si ricava la biiezione desiderata.

Come conseguenza del teorema di Schröder-Bernstein possiamo inoltre dimostrare che l'insieme $\mathcal{P}(\mathbb{N})$ di tutti i sottoinsiemi di $\mathbb{N}$ ha cardinalità c. Come notato sopra, basta dimostrare che $|\mathcal{P}(\mathbb{N}) \setminus \{\varnothing\}| = |(0,1]|$. Un esempio di trasformazione iniettiva è

Con questo abbiamo anche dimostrato un risultato annunciato in precedenza: *Ogni insieme infinito ha la stessa dimensione di un suo sottoinsieme proprio dato*

$$f : \mathcal{P}(\mathbb{N}) \setminus \{\varnothing\} \longrightarrow (0,1], \qquad A \longmapsto \sum_{i \in A} 10^{-i},$$

mentre

$$g : (0,1] \longrightarrow \mathcal{P}(\mathbb{N}) \setminus \{\varnothing\}, \qquad 0.b_1b_2b_3... \longmapsto \{b_i 10^i : i \in \mathbb{N}\}$$

definisce un'iniezione nell'altra direzione.

Finora conosciamo i numeri cardinali $0, 1, 2, \ldots, \aleph_0$, ed inoltre sappiamo che la cardinalità c di $\mathbb{R}$ è più grande di $\aleph_0$. Il passaggio da $\mathbb{Q}$ con $|\mathbb{Q}| = \aleph_0$ a $\mathbb{R}$ con $|\mathbb{R}| = c$ suggerisce immediatamente la seguente domanda:

> *il numero cardinale infinito $c = |\mathbb{R}|$ è quello che segue immediatamente $\aleph_0$?*

Ora, ovviamente, abbiamo il problema se *ci sia* un numero cardinale successivo, o in altre parole, se $\aleph_1$ abbia un significato. Ce l'ha — la dimostrazione è schematizzata nell'appendice di questo capitolo.

L'affermazione $c = \aleph_1$ divenne nota come l'*ipotesi del continuo*. La domanda se l'ipotesi del continuo fosse vera rappresentò per diversi decenni una delle sfide supreme di tutta la matematica. La risposta, alla fine fornita da Kurt Gödel e Paul Cohen, ci porta al limite del pensiero logico. Essi dimostrarono che l'affermazione $c = \aleph_1$ è *indipendente* dal sistema di assiomi di Zermelo-Fraenkel, nello stesso modo in cui l'assioma delle parallele è indipendente dagli altri assiomi della geometria euclidea. Vi sono modelli in cui vale $c = \aleph_1$ e vi sono altri modelli della teoria degli insiemi in cui $c \neq \aleph_1$.

Alla luce di questo fatto è interessante chiedersi se ci siano altre condizioni (ad esempio dall'analisi) che siano equivalenti all'ipotesi del continuo. In effetti, è naturale ricercare un esempio nell'analisi, dal momento che storicamente le prime applicazioni sostanziali della teoria degli insiemi di Cantor apparvero in quell'ambito, più precisamente nella teoria delle funzioni complesse. Di seguito vogliamo presentare un esempio di questo tipo e la

sua soluzione estremamente semplice ed elegante, fornita da Paul Erdős. Nel 1962, Wetzel pose la seguente domanda:

> *Sia $\{f_\alpha\}$ una famiglia di funzioni analitiche distinte a due a due sui numeri complessi tale che per ogni $z \in \mathbb{C}$ l'insieme dei valori $\{f_\alpha(z)\}$ sia al più numerabile (ovvero, finito); chiamiamo questa proprietà (P_0).*
> *Possiamo dedurre che la famiglia stessa è al più numerabile?*

Dopo brevissimo tempo Erdős dimostrò che, sorprendentemente, la risposta dipende dall'ipotesi del continuo.

Teorema 5. *Se $c > \aleph_1$, allora ogni famiglia $\{f_\alpha\}$ che soddisfi (P_0) è numerabile. Se, d'altro canto, $c = \aleph_1$, allora esiste una famiglia $\{f_\alpha\}$ di dimensione c soddisfacente la proprietà (P_0).*

Per la dimostrazione abbiamo bisogno di alcuni fatti basilari sui numeri cardinali e ordinali. Per i lettori non abituati a questi concetti, questo capitolo ha un'appendice dove sono raccolti tutti i risultati necessari.

■ **Dimostrazione del Teorema 5.** Si supponga dapprima che $c > \aleph_1$. Mostreremo che per ogni famiglia $\{f_\alpha\}$ di dimensione $\aleph_1$ di funzioni analitiche esiste un numero complesso z_0 tale che *tutti* gli $\aleph_1$ valori $f_\alpha(z_0)$ siano distinti. Di conseguenza, se una famiglia di funzioni soddisfa (P_0), allora deve essere numerabile.

Per vederlo, sfruttiamo le nostre conoscenze sui numeri ordinali. Innanzitutto, introduciamo un buon ordinamento nella famiglia $\{f_\alpha\}$ in base al numero ordinale iniziale ω_1 di $\aleph_1$. In base alla Proposizione 1 dell'appendice ciò significa che l'insieme degli indici spazia su tutti i numeri ordinali α inferiori a ω_1. Quindi mostriamo che l'insieme delle coppie (α, β), $\alpha < \beta < \omega_1$, ha dimensione $\aleph_1$. Poiché ogni $\beta < \omega_1$ è un numero ordinale numerabile, l'insieme delle coppie (α, β), $\alpha < \beta$, è numerabile per ogni β fisso. Prendendo l'unione su tutti gli $\aleph_1$ β, troviamo dalla Proposizione 6 dell'appendice che l'insieme di tutte le coppie (α, β), $\alpha < \beta$, ha dimensione $\aleph_1$.

Si consideri ora per ogni coppia $\alpha < \beta$ l'insieme

$$S(\alpha, \beta) \ = \ \{z \in \mathbb{C} : f_\alpha(z) = f_\beta(z)\}.$$

Affermiamo che ogni insieme $S(\alpha, \beta)$ è numerabile. Per verificarlo si considerino i dischi C_k di raggio $k = 1, 2, 3, \ldots$ centrati nell'origine del piano complesso. Se f_α e f_β hanno lo stesso valore su infiniti punti in un C_k, allora f_α e f_β sono identiche in base a un risultato ben noto sulle funzioni analitiche. Pertanto f_α e f_β hanno lo stesso valore solo su un numero finito di punti in ogni C_k, e dunque al massimo su un insieme numerabile di punti. Ora poniamo $S := \bigcup_{\alpha<\beta} S(\alpha, \beta)$. Di nuovo in base alla Proposizione 6, troviamo che S ha dimensione $\aleph_1$, poiché ogni insieme $S(\alpha, \beta)$ è numerabile. E qui arriva la sentenza: poiché, come sappiamo, $\mathbb{C}$ ha dimensione c

e c è maggiore di $\aleph_1$ per assunzione, esiste un numero complesso z_0 non in S e per questo z_0 tutti gli $\aleph_1$ valori $f_\alpha(z_0)$ sono distinti.

Supponiamo poi che $c = \aleph_1$. Si consideri l'insieme $D \subseteq \mathbb{C}$ dei numeri complessi $p + iq$ con parte reale razionale e parte immaginaria intera. Poiché per ogni p l'insieme $\{p + iq : q \in \mathbb{Q}\}$ è numerabile, troviamo che D è numerabile. Inoltre, D è un insieme *denso* in $\mathbb{C}$: ogni disco aperto nel piano complesso contiene un punto di D. Sia $\{z_\alpha : 0 \leq \alpha < \omega_1\}$ un buon ordinamento di $\mathbb{C}$. Costruiremo ora una famiglia $\{f_\beta : 0 \leq \beta < \omega_1\}$ di $\aleph_1$ funzioni analitiche distinte tale che

$$f_\beta(z_\alpha) \in D \text{ ogniqualvolta } \alpha < \beta. \tag{1}$$

Ognuna di tali famiglie soddisfa la condizione (P_0). In effetti, ogni punto $z \in \mathbb{C}$ ha un indice, diciamo $z = z_\alpha$. Ora, per ogni $\beta > \alpha$, i valori $\{f_\beta(z_\alpha)\}$ giacciono nell'insieme *numerabile* D. Poiché α è un numero ordinale numerabile, le funzioni f_β tali che $\beta \leq \alpha$ forniranno al massimo un insieme numerabile di ulteriori valori $f_\beta(z_\alpha)$, cosicché l'insieme di tutti i valori $\{f_\beta(z_\alpha)\}$ sarà analogamente al più numerabile. Pertanto, se possiamo costruire una famiglia $\{f_\beta\}$ che soddisfi (1), allora la seconda parte del teorema è dimostrata. La costruzione di $\{f_\beta\}$ è per induzione transfinita. Come f_0 possiamo prendere una qualunque funzione analitica, ad esempio $f_0 =$ costante. Supponiamo che f_β sia già stata costruita per ogni $\beta < \gamma$. Poiché γ è un ordinale numerabile, possiamo riordinare $\{f_\beta : 0 \leq \beta < \gamma\}$ in una successione $g_1, g_2, g_3, \ldots$. Lo stesso riordinamento di $\{z_\alpha : 0 \leq \alpha < \gamma\}$ dà una successione $w_1, w_2, w_3, \ldots$. Costruiremo ora una funzione f_γ che soddisfi per ogni n le condizioni

$$f_\gamma(w_n) \in D \qquad \text{e} \qquad f_\gamma(w_n) \neq g_n(w_n). \tag{2}$$

La seconda condizione assicurerà che tutte le funzioni f_γ $(0 \leq \gamma < \omega_1)$ siano distinte e la prima condizione è semplicemente (1), il che implica (P_0) secondo il nostro primo argomento. Si noti che la condizione $f_\gamma(w_n) \neq g_n(w_n)$ è una volta di più un argomento di diagonalizzazione. Per costruire f_γ, scriviamo

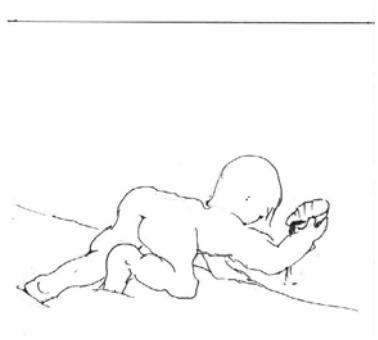
$$\begin{aligned} f_\gamma(z) \;&:=\; \varepsilon_0 + \varepsilon_1(z - w_1) + \varepsilon_2(z - w_1)(z - w_2) \\ &\quad + \varepsilon_3(z - w_1)(z - w_2)(z - w_3) + \ldots. \end{aligned}$$

Se γ è un ordinale finito, allora f_γ è un polinomio e pertanto una funzione analitica, e certamente possiamo scegliere dei numeri ε_i tali che (2) sia soddisfatta. Supponiamo ora che γ sia un ordinale numerabile, allora

$$f_\gamma(z) \;=\; \sum_{n=0}^{\infty} \varepsilon_n(z - w_1) \cdots (z - w_n). \tag{3}$$

Si noti che i valori di ε_m $(m \geq n)$ non hanno influenza sul valore $f_\gamma(w_n)$, pertanto possiamo scegliere ε_n passo per passo. Se la successione (ε_n) converge verso 0 abbastanza velocemente, allora (3) definisce una funzione

analitica. Infine, poiché D è un insieme denso, possiamo scegliere questa successione (ε_n) in modo tale che f_γ rispetti i requisiti di (2), e la dimostrazione è completa. □



"Una leggenda racconta di Sant'Agostino che, camminando sulla spiaggia e contemplando l'infinito, vide un bambino che tentava di svuotare l'oceano con una piccola conchiglia..."

Appendice: sui numeri cardinali e ordinali

Discutiamo dapprima la questione se per ogni numero cardinale ne esista uno successivo più grande. Per iniziare mostriamo che per ogni numero cardinale $\mathfrak{m}$ vi è sempre un numero cardinale $\mathfrak{n}$ più grande di $\mathfrak{m}$. A tal fine usiamo di nuovo una versione del metodo di diagonalizzazione di Cantor.

Sia M un insieme, allora affermiamo che l'insieme $\mathcal{P}(M)$ *di tutti i sottoinsiemi* di M ha dimensione maggiore di M. Facendo corrispondere $m \in M$ con $\{m\} \in \mathcal{P}(M)$, vediamo che M può essere trasformato biiettivamente in un sottoinsieme di $\mathcal{P}(M)$, il che implica $|M| \leq |\mathcal{P}(M)|$ per definizione. Resta da mostrare che $\mathcal{P}(M)$ *non può* essere trasformato biiettivamente in un sottoinsieme di M. Supponiamo, al contrario, che $\varphi : N \longrightarrow \mathcal{P}(M)$ sia una biiezione di $N \subseteq M$ su $\mathcal{P}(M)$. Si consideri il sottoinsieme $U \subseteq N$ di tutti gli elementi di N che *non* sono contenuti nella loro immagine attraverso φ, ovvero, $U = \{m \in N : m \notin \varphi(m)\}$. Poiché φ è una biiezione, esiste $u \in N$ tale che $\varphi(u) = U$. Ora, o $u \in U$ oppure $u \notin U$, ma entrambe le alternative sono impossibili! In effetti, se $u \in U$, allora $u \notin \varphi(u) = U$ in base alla definizione di U, e se $u \notin U = \varphi(u)$, allora $u \in U$, il che è una contraddizione.

Molto probabilmente, il lettore avrà già visto questo ragionamento. Si tratta del vecchio paradosso del barbiere: "Un barbiere è un uomo che rade tutti gli uomini che non si radono da sé. Il barbiere si rade da sé?".

Per addentrarci nella teoria introduciamo un altro grande concetto di Cantor, gli insiemi ordinati e i numeri ordinali. Un insieme M è *ordinato* rispetto a $<$ se la relazione $<$ è transitiva e se per ogni coppia di elementi distinti a e b di M abbiamo o $a < b$ o $b < a$. Ad esempio, possiamo ordinare $\mathbb{N}$ nel modo abituale secondo la grandezza, $\mathbb{N} = \{1, 2, 3, 4, \ldots\}$, ma, ovviamente, possiamo anche ordinare $\mathbb{N}$ nell'altro senso, $\mathbb{N} = \{\ldots, 4, 3, 2, 1\}$, oppure $\mathbb{N} = \{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$ elencando prima i numeri dispari e poi quelli pari.

Ecco il concetto di base. Un insieme ordinato M è detto *bene ordinato* se ogni sottoinsieme non vuoto di M ha un primo elemento. Pertanto il primo e terzo ordinamento di $\mathbb{N}$ sono buoni ordinamenti, ma il secondo non lo è. Il *teorema fondamentale del buon ordinamento*, implicato dagli assiomi (compreso l'assioma della scelta), afferma ora che *ogni* insieme M ammette un buon ordinamento. D'ora in poi, consideriamo soltanto insiemi dotati di un buon ordinamento.

Diciamo che due insiemi bene ordinati M ed N sono *simili* (o dello *stesso tipo di ordine*) se esiste una biiezione φ di M in N che rispetti l'ordine, ovvero, $m <_M n$ implica $\varphi(m) <_N \varphi(n)$. Si noti che ogni insieme ordinato simile ad un insieme bene ordinato è esso stesso bene ordinato.

Gli insiemi bene ordinati
$\mathbb{N} = \{1, 2, 3, \ldots\}$ e
$\mathbb{N} = \{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$
non sono simili: il primo ordinamento ha solo un elemento senza un predecessore immediato, mentre il secondo ne ha due

La similitudine è ovviamente una relazione di equivalenza e possiamo pertanto parlare di un *numero ordinale* α che appartenga ad una classe di insiemi simili. Per insiemi finiti, gli elementi di ogni coppia di insiemi ordinati qualunque hanno dei buoni ordinamenti simili ed usiamo di nuovo il numero ordinale n per la classe degli n-insiemi. Si noti che, per definizione, due insiemi simili hanno la stessa cardinalità. Pertanto ha senso parlare di *cardinalità* $|\alpha|$ di un numero ordinale α. Si noti inoltre che ogni sottoinsieme di un insieme bene ordinato è anch'esso bene ordinato rispetto all'ordinamento indotto.

Come abbiamo fatto per i numeri cardinali, paragoniamo ora i numeri ordinali. Sia M un insieme bene ordinato, $m \in M$, allora $M_m = \{x \in M : x < m\}$ è chiamato il *segmento* (*iniziale*) di M determinato da m; N è un segmento di M se $N = M_m$ per un m dato. Pertanto, in particolare, M_m è l'insieme vuoto quando m è il primo elemento di M. Ora siano μ e ν i numeri ordinali degli insiemi ben ordinati M ed N. Diciamo che μ è *più piccolo* di ν, $\mu < \nu$, se M è simile a un segmento di N. Di nuovo, abbiamo la legge transitiva che afferma che $\mu < \nu$, $\nu < \pi$ implica $\mu < \pi$, poiché sotto una trasformazione di similitudine un segmento è trasformato in un altro segmento.

Il numero ordinale di $\{1, 2, 3, \ldots\}$ è inferiore al numero ordinale di $\{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$

Chiaramente, per insiemi finiti, $m < n$ ha il significato consueto. Indichiamo con ω il numero ordinale di $\mathbb{N} = \{1, 2, 3, 4, \ldots\}$ munito dell'ordinamento naturale. Considerando il segmento $\mathbb{N}_{n+1}$ troviamo $n < \omega$ per ogni n finito. Ora vediamo che $\omega \leq \alpha$ vale per ogni numero ordinale infinito α. In effetti, se l'insieme infinito bene ordinato M ha come numero ordinale α, allora M contiene come primo elemento m_1, $M\backslash\{m_1\}$ contiene come primo elemento m_2, $M\backslash\{m_1, m_2\}$ contiene come primo elemento m_3. Continuando in questo modo, produciamo la successione $m_1 < m_2 < m_3 < \ldots$ in M. Se $M = \{m_1, m_2, m_3, \ldots\}$, allora M è simile a $\mathbb{N}$, e dunque $\alpha = \omega$. Se, d'altro canto, $M\backslash\{m_1, m_2, \ldots\}$ è non vuoto, allora contiene come primo elemento m, e concludiamo che $\mathbb{N}$ è simile al segmento M_m, ovvero, $\omega < \alpha$ per definizione.

Affermiamo ora (senza dimostrazioni, peraltro non difficili) tre risultati basilari sui numeri ordinali. Il primo afferma che ogni numero ordinale μ ha un insieme rappresentante bene ordinato "standard", W_μ.

Proposizione 1. *Sia μ un numero ordinale e si indichi con W_μ l'insieme dei numeri ordinali più piccoli di μ. Allora vale quanto segue:*

(i) *Gli elementi di W_μ sono confrontabili a due a due.*

(ii) *Se ordiniamo W_μ naturalmente, allora W_μ è bene ordinato ed ha come numero ordinale μ.*

Proposizione 2. *Ogni coppia di numeri ordinali μ e ν soddisfa esattamente una delle relazioni $\mu < \nu$, $\mu = \nu$, o $\mu > \nu$.*

Proposizione 3. *Ogni insieme di numeri ordinali (ordinati secondo la magnitudine) è bene ordinato.*

Dopo questa incursione nei numeri ordinali torniamo ai numeri cardinali. Sia $\mathbf{m}$ un numero cardinale e si indichi con $O_\mathbf{m}$ l'insieme di tutti i numeri

ordinali μ tali che $|\mu| = \mathfrak{m}$. Secondo la Proposizione 3 esiste un numero ordinale *minimo* $\omega_{\mathfrak{m}}$ in $O_{\mathfrak{m}}$, che chiamiamo *numero ordinale iniziale* di $\mathfrak{m}$. Ad esempio, ω è il numero ordinale iniziale di $\aleph_0$.

Con queste premesse possiamo ora dimostrare un risultato basilare per questo capitolo.

Proposizione 4. *Ogni numero cardinale $\mathfrak{m}$ ha un successivo.*

■ **Dimostrazione.** Sappiamo già che esiste un numero cardinale più grande $\mathfrak{n}$. Si consideri ora l'insieme $\mathcal{K}$ di tutti i numeri cardinali maggiori di $\mathfrak{m}$ e grandi al massimo quanto $\mathfrak{n}$. Associamo ad ogni $\mathfrak{p} \in \mathcal{K}$ il proprio numero ordinale iniziale $\omega_{\mathfrak{p}}$. Fra questi numeri iniziali ve ne è uno più piccolo (come si ottiene dalla Proposizione 3); il numero cardinale corrispondente è quindi il minimo di $\mathcal{K}$ e dunque è il numero cardinale successivo a $\mathfrak{m}$ che si stava cercando. □

Proposizione 5. *Sia $\mathfrak{m}$ la cardinalità dell'insieme infinito M e sia M bene ordinato in base al numero ordinale iniziale $\omega_{\mathfrak{m}}$. Allora M non ha un ultimo elemento.*

■ **Dimostrazione.** Effettivamente, se M avesse un ultimo elemento m, allora il segmento M_m avrebbe un numero ordinale $\mu < \omega_{\mathfrak{m}}$ tale che $|\mu| = \mathfrak{m}$, contraddicendo la definizione di $\omega_{\mathfrak{m}}$. □

Per concludere abbiamo bisogno di rafforzare considerabilmente il risultato secondo cui l'unione numerabile di insiemi numerabili è ancora numerabile. Nella proposizione che segue consideriamo famiglie *arbitrarie* di insiemi numerabili.

Proposizione 6. *Supponiamo che $\{A_\alpha\}$ sia una famiglia di dimensione $\mathfrak{m}$ di insiemi numerabili A_α, dove $\mathfrak{m}$ è un cardinale infinito. Allora l'unione $\bigcup_\alpha A_\alpha$ ha al massimo dimensione $\mathfrak{m}$.*

■ **Dimostrazione.** Possiamo supporre che gli insiemi A_α siano disgiunti a due a due, dal momento che questo può solo aumentare la dimensione dell'unione. Sia M, con $|M| = \mathfrak{m}$, l'insieme degli indici e se ne faccia un buon ordinamento in base al numero ordinale iniziale $\omega_{\mathfrak{m}}$. Sostituiamo ora ogni $\alpha \in M$ con un insieme numerabile $B_\alpha = \{b_{\alpha 1} = \alpha, b_{\alpha 2}, b_{\alpha 3}, \ldots\}$, ordinato in base a ω, e chiamiamo il nuovo insieme $\widetilde{M}$. Allora $\widetilde{M}$ è nuovamente bene ordinato se si pone $b_{\alpha i} < b_{\beta j}$ per $\alpha < \beta$ e $b_{\alpha i} < b_{\alpha j}$ per $i < j$. Sia $\widetilde{\mu}$ il numero ordinale di $\widetilde{M}$. Poiché M è un sottoinsieme di $\widetilde{M}$, abbiamo $\mu \leq \widetilde{\mu}$ in base ad un'argomentazione precedente. Se $\mu = \widetilde{\mu}$, allora M è simile a $\widetilde{M}$ e se $\mu < \widetilde{\mu}$, allora M è simile ad un segmento di $\widetilde{M}$. Ora, poiché l'ordinamento $\omega_{\mathfrak{m}}$ di M non ha ultimi elementi (Proposizione 5), vediamo che M è in entrambi i casi simile all'unione degli insiemi numerabili B_β e pertanto della stessa cardinalità.

Il resto è semplice. Sia $\varphi : \bigcup B_\beta \longrightarrow M$ una biiezione, e si supponga che $\varphi(B_\beta) = \{\alpha_1, \alpha_2, \alpha_3, \ldots\}$. Si sostituisca ogni α_i con A_{α_i} e si consideri l'unione $\bigcup A_{\alpha_i}$. Poiché $\bigcup A_{\alpha_i}$ è l'unione di un numero *numerabile* di

insiemi numerabili (e pertanto è numerabile) vediamo che B_β ha la stessa dimensione di $\bigcup A_{\alpha_i}$. In altre parole, vi è biiezione da B_β a $\bigcup A_{\alpha_i}$ per ogni β e dunque una biiezione ψ da $\bigcup B_\beta$ su $\bigcup A_\alpha$. Ma ora $\psi\varphi^{-1}$ dà la biiezione desiderata di M su $\bigcup A_\alpha$ e dunque $|\bigcup A_\alpha| = \mathfrak{m}$. □

Bibliografia

[1] L. E. J. BROUWER: *Beweis der Invarianz der Dimensionszahl,* Math. Annalen **70** (1911), 161-165.

[2] N. CALKIN & H. WILF: *Recounting the rationals,* Amer. Math. Monthly **107** (2000), 360-363.

[3] P. COHEN: *Set Theory and the Continuum Hypothesis,* W. A. Benjamin, New York 1966.

[4] P. ERDŐS: *An interpolation problem associated with the continuum hypothesis,* Michigan Math. J. **11** (1964), 9-10.

[5] E. KAMKE: *Theory of Sets,* Dover Books 1950.

[6] M. A. STERN: *Ueber eine zahlentheoretische Funktion,* Journal für die reine und angewandte Mathematik **55** (1858), 193-220.

"Infinitamente molti più cardinali"

Elogio delle disuguaglianze

Capitolo 17

L'analisi abbonda di disuguaglianze, come testimonia ad esempio la celebre opera "Disuguaglianze" di Hardy, Littlewood e Pólya. Consideriamo due delle disuguaglianze più basilari, ciascuna con due applicazioni, e seguiamo lo stesso George Pólya, che fu un campione in materia di *Book Proof*, per scoprire quali dimostrazioni consideri più appropriate.

La nostra prima disuguaglianza è attribuita ora a Cauchy, ora a Schwarz e/o a Buniakowski:

Teorema I (Disuguaglianza di Cauchy-Schwarz)
Sia $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ un prodotto interno definito su uno spazio vettoriale reale V (con norma $|\boldsymbol{a}|^2 := \langle \boldsymbol{a}, \boldsymbol{a} \rangle$). Allora

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \ \leq \ |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$$

per tutti i vettori $\boldsymbol{a}, \boldsymbol{b} \in V$, con uguaglianza se e solo se $\boldsymbol{a}$ e $\boldsymbol{b}$ sono linearmente dipendenti.

■ **Dimostrazione.** La seguente (folcloristica) dimostrazione è probabilmente la più breve. Si consideri la funzione quadratica

$$|x\boldsymbol{a} + \boldsymbol{b}|^2 \ = \ x^2 |\boldsymbol{a}|^2 + 2x \langle \boldsymbol{a}, \boldsymbol{b} \rangle + |\boldsymbol{b}|^2$$

nella variabile x. Possiamo supporre $\boldsymbol{a} \neq \boldsymbol{0}$. Se $\boldsymbol{b} = \lambda \boldsymbol{a}$, allora chiaramente $\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 = |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$. Se, d'altro canto, $\boldsymbol{a}$ e $\boldsymbol{b}$ sono linearmente indipendenti, allora $|x\boldsymbol{a} + \boldsymbol{b}|^2 > 0$ per ogni x e pertanto il discriminante $\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 - |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$ è minore di 0. □

Il nostro secondo esempio è la *disuguaglianza della media armonica, geometrica ed aritmetica*:

Teorema II (Media armonica, geometrica ed aritmetica)
Siano $a_1, \ldots, a_n$ numeri reali positivi, allora

$$\frac{n}{\frac{1}{a_1} + \ldots + \frac{1}{a_n}} \ \leq \ \sqrt[n]{a_1 a_2 \ldots a_n} \ \leq \ \frac{a_1 + \ldots + a_n}{n}$$

con uguaglianza in entrambi i casi se e solo se tutti gli a_i sono uguali.

■ **Dimostrazione.** La splendida e inconsueta dimostrazione per induzione che segue è attribuita a Cauchy (si veda [7]). Sia $P(n)$ la proprietà espressa dalla seconda disuguaglianza, che scriviamo nella forma

$$a_1 a_2 \ldots a_n \ \leq \ \Big(\frac{a_1 + \ldots + a_n}{n} \Big)^n .$$

Per $n = 2$, abbiamo $a_1 a_2 \le (\frac{a_1+a_2}{2})^2 \iff (a_1 - a_2)^2 \ge 0$, il che è vero. Ora procediamo con i due passi seguenti:

(A) $P(n) \Longrightarrow P(n-1)$

(B) $P(n)$ e $P(2) \Longrightarrow P(2n)$

i quali evidentemente implicheranno il risultato completo.

Per dimostrare **(A)**, si ponga $A := \sum_{k=1}^{n-1} \frac{a_k}{n-1}$, allora

$$\Big(\prod_{k=1}^{n-1} a_k\Big) A \overset{P(n)}{\le} \left(\frac{\sum_{k=1}^{n-1} a_k \;+\; A}{n}\right)^n \;=\; \Big(\frac{(n-1)A + A}{n}\Big)^n \;=\; A^n$$

e pertanto $\displaystyle\prod_{k=1}^{n-1} a_k \;\le\; A^{n-1} \;=\; \left(\frac{\sum_{k=1}^{n-1} a_k}{n-1}\right)^{n-1}$.

Per **(B)**, osserviamo che

$$\prod_{k=1}^{2n} a_k = \Big(\prod_{k=1}^{n} a_k\Big)\Big(\prod_{k=n+1}^{2n} a_k\Big) \overset{P(n)}{\le} \Big(\sum_{k=1}^{n} \frac{a_k}{n}\Big)^n \Big(\sum_{k=n+1}^{2n} \frac{a_k}{n}\Big)^n$$
$$\overset{P(2)}{\le} \left(\frac{\sum_{k=1}^{2n} \frac{a_k}{n}}{2}\right)^{2n} = \left(\frac{\sum_{k=1}^{2n} a_k}{2n}\right)^{2n}.$$

La condizione per l'uguaglianza si deriva altrettanto semplicemente.

La disuguaglianza di sinistra, tra la media armonica e quella geometrica, segue ora considerando $\frac{1}{a_1}, \ldots, \frac{1}{a_n}$. □

■ **Un'altra dimostrazione.** Fra le diverse altre dimostrazioni della disuguaglianza fra la media aritmetica e quella geometrica (la monografia [2] ne elenca oltre 50), ne isoliamo una recente particolarmente notevole ad opera di Alzer. In effetti, questa dimostrazione fornisce la disuguaglianza più forte

$$a_1^{p_1} a_2^{p_2} \ldots a_n^{p_n} \;\le\; p_1 a_1 + p_2 a_2 + \ldots + p_n a_n$$

per tutti i numeri positivi $a_1, \ldots, a_n, p_1, \ldots, p_n$ con $\sum_{i=1}^n p_i = 1$. Indichiamo l'espressione di sinistra con G e quella di destra con A. Possiamo supporre che $a_1 \le \ldots \le a_n$. Chiaramente $a_1 \le G \le a_n$, dunque deve esistere un k tale che $a_k \le G \le a_{k+1}$. Ne segue che

$$\sum_{i=1}^{k} p_i \int_{a_i}^{G} \Big(\frac{1}{t} - \frac{1}{G}\Big)\, dt \;+\; \sum_{i=k+1}^{n} p_i \int_{G}^{a_i} \Big(\frac{1}{G} - \frac{1}{t}\Big)\, dt \;\ge\; 0 \qquad (1)$$

poiché tutti gli integrandi sono ≥ 0. Riscrivendo (1) otteniamo

$$\sum_{i=1}^{n} p_i \int_{G}^{a_i} \frac{1}{G}\, dt \;\ge\; \sum_{i=1}^{n} p_i \int_{G}^{a_i} \frac{1}{t}\, dt$$

dove il membro di sinistra è uguale a

$$\sum_{i=1}^{n} p_i \frac{a_i - G}{G} \;=\; \frac{A}{G} - 1,$$

mentre quello di destra è

$$\sum_{i=1}^{n} p_i(\log a_i - \log G) = \log \prod_{i=1}^{n} a_i^{p_i} - \log G = 0.$$

Concludiamo che $\frac{A}{G} - 1 \geq 0$ ovvero $A \geq G$. Nel caso dell'uguaglianza, tutti gli integrali in (1) devono essere 0, il che implica $a_1 = \ldots = a_n = G$. □

La nostra prima applicazione è uno splendido risultato di Laguerre (si veda [7]) riguardante la localizzazione delle radici dei polinomi.

Teorema 1. *Se tutte le radici del polinomio* $x^n + a_{n-1}x^{n-1} + \ldots + a_0$ *sono reali, allora sono contenute nell'intervallo di estremi*

$$-\frac{a_{n-1}}{n} \pm \frac{n-1}{n}\sqrt{a_{n-1}^2 - \frac{2n}{n-1}a_{n-2}}\,.$$

■ **Dimostrazione.** Sia y una delle radici e siano $y_1, \ldots, y_{n-1}$ le altre. Allora il polinomio si scrive $(x - y)(x - y_1)\cdots(x - y_{n-1})$. Pertanto confrontando i coefficienti

$$\begin{aligned} -a_{n-1} &= y + y_1 + \ldots + y_{n-1}, \\ a_{n-2} &= y(y_1 + \ldots + y_{n-1}) + \sum_{i<j} y_i y_j, \end{aligned}$$

e dunque

$$a_{n-1}^2 - 2a_{n-2} - y^2 \;=\; \sum_{i=1}^{n-1} y_i^2.$$

Secondo la disuguaglianza di Cauchy applicata a $(y_1, \ldots, y_{n-1})$ e $(1, \ldots, 1)$,

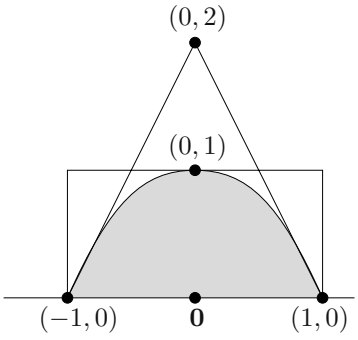
$$\begin{aligned} (a_{n-1} + y)^2 &= (y_1 + y_2 + \ldots + y_{n-1})^2 \\ &\leq (n-1)\sum_{i=1}^{n-1} y_i^2 \;=\; (n-1)(a_{n-1}^2 - 2a_{n-2} - y^2), \end{aligned}$$

ovvero

$$y^2 + \frac{2a_{n-1}}{n}y + \frac{2(n-1)}{n}a_{n-2} - \frac{n-2}{n}a_{n-1}^2 \leq 0.$$

Pertanto y (e dunque tutti gli y_i) giace tra le due radici della funzione quadratica e tali radici sono i nostri limiti. □

Per la nostra seconda applicazione iniziamo da una ben nota proprietà elementare delle parabole. Si consideri la parabola descritta da $f(x) = 1 - x^2$

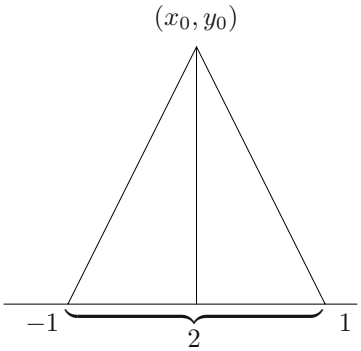


tra $x = -1$ e $x = 1$. Associamo a $f(x)$ il *triangolo tangente* ed il *rettangolo tangente* come in figura.

Troviamo che l'area ombreggiata $A = \int_{-1}^{1}(1-x^2)dx$ è uguale a $\frac{4}{3}$ e le aree T e R del triangolo e del rettangolo sono entrambe uguali a 2. Pertanto $\frac{T}{A} = \frac{3}{2}$ e $\frac{R}{A} = \frac{3}{2}$. In una splendida pubblicazione, Paul Erdős e Tibor Gallai si chiesero cosa accade quando $f(x)$ è un polinomio reale arbitrario di grado n con $f(x) > 0$ per $-1 < x < 1$, e $f(-1) = f(1) = 0$. L'area A vale dunque $\int_{-1}^{1} f(x)dx$. Supponiamo che $f(x)$ assuma in $(-1, 1)$ il suo valore massimo in corrispondenza di b, allora $R = 2f(b)$. Calcolando le tangenti in -1 ed 1, appare evidente (si veda il riquadro) che

$$T \;=\; \frac{2f'(1)f'(-1)}{f'(1) - f'(-1)}, \tag{2}$$

e $T = 0$ per $f'(1) = f'(-1) = 0$.



Il triangolo tangente

L'area T del triangolo tangente è esattamente y_0, dove (x_0, y_0) è il punto di intersezione delle due tangenti. Le equazioni di queste tangenti sono $y = f'(-1)(x + 1)$ e $y = f'(1)(x - 1)$, pertanto

$$x_0 \;=\; \frac{f'(1) + f'(-1)}{f'(1) - f'(-1)},$$

e dunque

$$y_0 \;=\; f'(1)\Big(\frac{f'(1) + f'(-1)}{f'(1) - f'(-1)} - 1\Big) \;=\; 2\frac{f'(1)f'(-1)}{f'(1) - f'(-1)}.$$

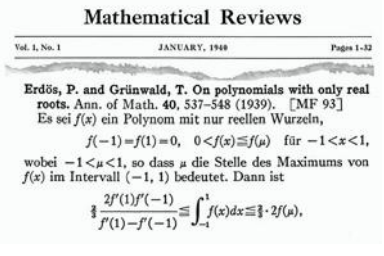
In generale, non vi sono limiti non banali per $\frac{T}{A}$ e $\frac{R}{A}$. Per constatarlo, si prenda $f(x) = 1 - x^{2n}$. Allora $T = 2n$, $A = \frac{4n}{2n+1}$ e pertanto $\frac{T}{A} > n$. Analogamente, $R = 2$ e $\frac{R}{A} = \frac{2n+1}{2n}$, che tende ad 1 quando n tende all'infinito.

Tuttavia, come mostrarono Erdős e Gallai, per polinomi che hanno solo radici reali, tali limiti esistono effettivamente.

Teorema 2. *Sia $f(x)$ un polinomio reale di grado $n \geq 2$ con solo radici reali, tale che $f(x) > 0$ per $-1 < x < 1$ e $f(-1) = f(1) = 0$. Allora*

$$\frac{2}{3}T \;\leq\; A \;\leq\; \frac{2}{3}R,$$

e l'uguaglianza vale in entrambi i casi solo per $n = 2$.


Mathematical Reviews

Vol. 1, No. 1 JANUARY, 1940 Pages 1-32

Erdös, P. and Grünwald, T. On polynomials with only real roots. Ann. of Math. **40**, 537–548 (1939). [MF 93]
Es sei $f(x)$ ein Polynom mit nur reellen Wurzeln,

$$f(-1)=f(1)=0, \quad 0<f(x)\leqq f(\mu) \quad \text{für} \quad -1<x<1,$$

wobei $-1<\mu<1$, so dass μ die Stelle des Maximums von $f(x)$ im Intervall $(-1, 1)$ bedeutet. Dann ist

$$\tfrac{2}{3}\frac{2f'(1)f'(-1)}{f'(1)-f'(-1)}\leqq\int_{-1}^{1}f(x)dx\leqq\tfrac{2}{3}\cdot 2f(\mu),$$

Erdős e Gallai stabilirono questo risultato con un'intricata dimostrazione per induzione. Nella recensione al loro articolo, che apparve sulla prima pagina del primo numero del Mathematical Reviews nel 1940, George

Pólya spiegò come la prima disuguaglianza potesse anche essere dimostrata dalla disuguaglianza tra la media aritmetica e la media geometrica — uno splendido esempio di commento costruttivo oltre che una *Book Proof*.

■ **Dimostrazione di $\frac{2}{3}T \le A$.** Poiché $f(x)$ ha solo radici reali, e nessuna di esse nell'intervallo aperto $(-1,1)$, si può scrivere — a parte un fattore positivo costante che si elimina alla fine — nella forma

$$f(x) \;=\; (1-x^2)\prod_i(\alpha_i - x)\prod_j(\beta_j + x) \tag{3}$$

con $\alpha_i \ge 1$, $\beta_j \ge 1$. Pertanto

$$A \;=\; \int_{-1}^{1}(1-x^2)\prod_i(\alpha_i - x)\prod_j(\beta_j + x)dx.$$

Effettuando la sostituzione $x \longmapsto -x$, troviamo anche che

$$A \;=\; \int_{-1}^{1}(1-x^2)\prod_i(\alpha_i + x)\prod_j(\beta_j - x)dx,$$

e pertanto secondo la disuguaglianza della media aritmetica e geometrica (si noti che tutti i fattori sono ≥ 0)

$$\begin{aligned} A \;&=\; \int_{-1}^{1}\frac{1}{2}\Big[\;(1-x^2)\prod_i(\alpha_i - x)\prod_j(\beta_j + x)\;+ \\ &\qquad\qquad (1-x^2)\prod_i(\alpha_i + x)\prod_j(\beta_j - x)\;\Big]dx \\ &\ge\; \int_{-1}^{1}(1-x^2)\Big(\prod_i(\alpha_i^2 - x^2)\prod_j(\beta_j^2 - x^2)\Big)^{1/2}dx \\ &\ge\; \int_{-1}^{1}(1-x^2)\Big(\prod_i(\alpha_i^2 - 1)\prod_j(\beta_j^2 - 1)\Big)^{1/2}dx \\ &=\; \frac{4}{3}\Big(\prod_i(\alpha_i^2 - 1)\prod_j(\beta_j^2 - 1)\Big)^{1/2}. \end{aligned}$$

Calcoliamo $f'(1)$ e $f'(-1)$. (Possiamo supporre che $f'(-1), f'(1) \ne 0$, poiché altrimenti $T = 0$ e la disuguaglianza $\frac{2}{3}T \le A$ diventa banale.) In base a (3) vediamo

$$f'(1) \;=\; -2\prod_i(\alpha_i - 1)\prod_j(\beta_j + 1)\,,$$

ed analogamente

$$f'(-1) \;=\; 2\prod_i(\alpha_i + 1)\prod_j(\beta_j - 1)\,.$$

Pertanto concludiamo

$$A \;\geq\; \frac{2}{3}(-f'(1)f'(-1))^{1/2}.$$

Applicando ora la disuguaglianza della media armonica e geometrica a $-f'(1)$ e $f'(1)$, arriviamo mediante (2) alla conclusione

$$A \;\geq\; \frac{2}{3}\frac{2}{\frac{1}{-f'(1)}+\frac{1}{f'(-1)}} \;=\; \frac{4}{3}\frac{f'(1)f'(-1)}{f'(1)-f'(-1)} \;=\; \frac{2}{3}T,$$

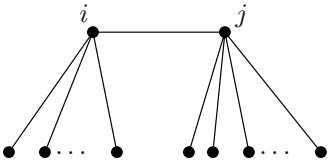
che è quanto volevamo dimostrare. Analizzando il caso dell'uguaglianza in tutte le nostre disuguaglianze il lettore potrà facilmente produrre l'ultima tesi del teorema. □

Il lettore è invitato a cercare una dimostrazione altrettanto ispirata della seconda disuguaglianza nel Teorema 2.

L'analisi è fatta di disuguaglianze dopo tutto, ma ecco un esempio tratto dalla teoria dei grafi dove l'uso delle disuguaglianze arriva in modo alquanto inatteso. Nel Capitolo 32 discuteremo il teorema di Turán. Nel caso più semplice esso prende la forma seguente:

Teorema 3. *Si supponga che G sia un grafo su n vertici senza triangoli. Allora G ha al massimo $\frac{n^2}{4}$ spigoli e l'uguaglianza vale solo quando n è pari e G è il grafo completo bipartito $K_{n/2,n/2}$.*

■ **Prima dimostrazione.** Questa dimostrazione, che usa la disuguaglianza di Cauchy, si deve a Mantel. Siano $V = \{1,\dots,n\}$ l'insieme dei vertici e E l'insieme degli spigoli di G. Indichiamo con d_i il grado di i, pertanto $\sum_{i\in V} d_i = 2|E|$ (si veda a pagina 161 nel capitolo sulla conta doppia). Supponiamo che ij sia uno spigolo. Essendo G senza triangoli, troviamo $d_i + d_j \leq n$ poiché nessun vertice è vicino di i e j al contempo.



Ne segue che

$$\sum_{ij\in E}(d_i+d_j) \;\leq\; n|E|.$$

Si noti che d_i appare esattamente d_i volte nella somma, quindi abbiamo

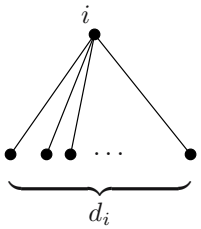
$$n|E| \;\geq\; \sum_{ij\in E}(d_i+d_j) \;=\; \sum_{i\in V} d_i^2,$$

e dunque con la disuguaglianza di Cauchy applicata ai vettori $(d_1,\dots,d_n)$ e $(1,\dots,1)$,

$$n|E| \;\geq\; \sum_{i\in V} d_i^2 \;\geq\; \frac{(\sum d_i)^2}{n} \;=\; \frac{4|E|^2}{n},$$

da cui segue il risultato. Nel caso dell'uguaglianza troviamo $d_i = d_j$ per ogni i, j ed inoltre $d_i = \frac{n}{2}$ (poiché $d_i + d_j = n$). Dal momento che G è senza triangoli, ne deriva immediatamente $G = K_{n/2,n/2}$. □

■ **Seconda dimostrazione.** La seguente dimostrazione del Teorema 3, che fa uso della disuguaglianza delle medie aritmetica e geometrica, è una *Book Proof* folcloristica. Sia α la dimensione di uno degli insiemi indipendenti più grandi, A, e si ponga $\beta = n - \alpha$. Poiché G è senza triangoli, i vicini di un vertice i formano un insieme indipendente e deduciamo $d_i \leq \alpha$ per ogni i.



L'insieme $B = V \setminus A$ di dimensione β incontra ogni spigolo di G. Contando gli spigoli di G in base ai loro vertici in B, otteniamo $|E| \leq \sum_{i \in B} d_i$. La disuguaglianza fra la media aritmetica e geometrica ora dà

$$|E| \;\leq\; \sum_{i \in B} d_i \;\leq\; \alpha\beta \;\leq\; \Big(\frac{\alpha+\beta}{2}\Big)^2 \;=\; \frac{n^2}{4},$$

e ancora una volta il caso dell'uguaglianza si tratta con facilità. □

Bibliografia

[1] H. ALZER: *A proof of the arithmetic mean-geometric mean inequality,* Amer. Math. Monthly **103** (1996), 585.

[2] P. S. BULLEN, D. S. MITRINOVICS & P. M. VASIĆ: *Means and their Inequalities,* Reidel, Dordrecht 1988.

[3] P. ERDŐS & T. GRÜNWALD: *On polynomials with only real roots,* Annals Math. **40** (1939), 537-548.

[4] G. H. HARDY, J. E. LITTLEWOOD & G. PÓLYA: *Inequalities,* Cambridge University Press, Cambridge 1952.

[5] W. MANTEL: *Problem 28,* Wiskundige Opgaven **10** (1906), 60-61.

[6] G. PÓLYA: *Review of* [3], Mathematical Reviews **1** (1940), 1.

[7] G. PÓLYA & G. SZEGŐ: *Problems and Theorems in Analysis, Vol. I,* Springer-Verlag, Berlin Heidelberg New York 1972/78; Reprint 1998.

Un teorema di Pólya sui polinomi

Capitolo 18

Tra i molti contributi di George Pólya all'analisi, quello riportato qui di seguito è sempre stato il preferito da Erdős, sia per il suo risultato sorprendente sia per la bellezza della dimostrazione. Supponiamo che

$$f(z) \;=\; z^n + b_{n-1}z^{n-1} + \ldots + b_0$$

sia un polinomio complesso di grado $n \geq 1$ monico (ovvero con coefficiente direttivo 1). Si associ a $f(z)$ l'insieme

$$\mathcal{C} \;:=\; \{z \in \mathbb{C} : |f(z)| \leq 2\},$$

ovvero, $\mathcal{C}$ è l'insieme dei punti trasformati da f nel cerchio di raggio 2 intorno all'origine nel piano complesso. Pertanto per $n = 1$ il dominio $\mathcal{C}$ è semplicemente un disco circolare di diametro 4.

Con un argomento sbalorditivamente semplice, Pólya rivelò la splendida proprietà di tale insieme $\mathcal{C}$:

> *Si prenda una retta L nel piano complesso e si consideri la proiezione ortogonale $\mathcal{C}_L$ dell'insieme $\mathcal{C}$ su L. Allora la lunghezza totale di tale proiezione non supera mai* 4.

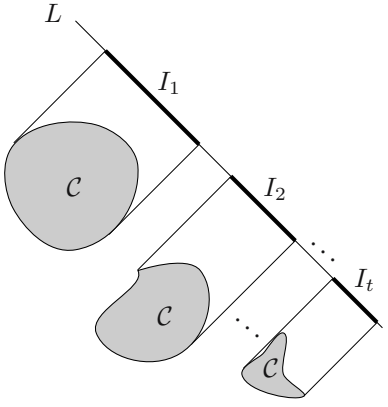
George Pólya

Che cosa significa che la lunghezza totale della proiezione $\mathcal{C}_L$ è al massimo 4? Vedremo che $\mathcal{C}_L$ *è* un'unione finita di intervalli disgiunti $I_1, \ldots, I_t$ e la condizione significa che $\ell(I_1)+\ldots+\ell(I_t) \leq 4$ dove $\ell(I_j)$ è la lunghezza abituale di un intervallo.

Ruotando il piano vediamo che basta considerare il caso in cui L sia l'asse reale del piano complesso. Avendo presente questi commenti, enunciamo il risultato di Pólya.

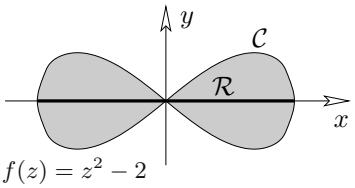
Teorema 1. *Sia $f(z)$ un polinomio complesso monico di grado almeno* 1. *Si ponga $\mathcal{C} = \{z \in \mathbb{C} : |f(z)| \leq 2\}$ e sia $\mathcal{R}$ la proiezione ortogonale di $\mathcal{C}$ sull'asse reale. Allora vi sono degli intervalli $I_1, \ldots, I_t$ sulla retta reale che insieme ricoprono $\mathcal{R}$ e soddisfano*

$$\ell(I_1) + \ldots + \ell(I_t) \;\leq\; 4.$$

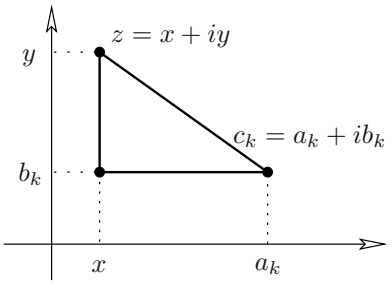


Chiaramente il limite 4 nel teorema è raggiunto per $n = 1$. Per avere una percezione più chiara del problema osserviamo il polinomio $f(z) = z^2 - 2$, che tocca anch'esso il limite 4. Se $z = x+iy$ è un numero complesso, allora x è la sua proiezione ortogonale sulla retta reale. Pertanto

$$\mathcal{R} = \{x \in \mathbb{R} : x + iy \in \mathcal{C} \text{ per qualche } y\}.$$



$f(z) = z^2 - 2$



Il lettore può facilmente dimostrare che per $f(z) = z^2-2$ abbiamo $x+iy \in \mathcal{C}$ se e solo se

$$(x^2 + y^2)^2 \ \leq \ 4(x^2 - y^2).$$

Ne segue che $x^4 \leq (x^2 + y^2)^2 \leq 4x^2$ e pertanto $x^2 \leq 4$ ovvero, $|x| \leq 2$. D'altro canto, ogni $z = x \in \mathbb{R}$ tale che $|x| \leq 2$ soddisfa $|z^2 - 2| \leq 2$ e troviamo che $\mathcal{R}$ è precisamente l'intervallo $[-2, 2]$ di lunghezza 4.

Come primo passo verso la dimostrazione scriviamo $f(z) = (z-c_1)\cdots(z-c_n)$ con $c_k = a_k + ib_k$ e consideriamo il polinomio *reale* $p(x) = (x - a_1)\cdots(x - a_n)$. Sia $z = x + iy \in \mathcal{C}$, allora grazie al teorema di Pitagora

$$|x - a_k|^2 + |y - b_k|^2 = |z - c_k|^2$$

e pertanto $|x - a_k| \leq |z - c_k|$ per ogni k, ovvero,

$$|p(x)| \ = \ |x - a_1|\cdots|x - a_n| \ \leq \ |z - c_1|\cdots|z - c_n| \ = \ |f(z)| \leq 2.$$

Troviamo dunque che $\mathcal{R}$ è incluso nell'insieme $\mathcal{P} = \{x \in \mathbb{R} : |p(x)| \leq 2\}$ e, se possiamo dimostrare che quest'ultimo insieme è coperto da intervalli di lunghezza totale massima 4, allora abbiamo finito. Analogamente, il nostro Teorema 1 principale sarà una conseguenza del risultato seguente.

Teorema 2. *Sia $p(x)$ un polinomio reale di grado $n \geq 1$ monico e con tutte radici reali. Allora l'insieme $\mathcal{P} = \{x \in \mathbb{R} : |p(x)| \leq 2\}$ si può ricoprire con intervalli di lunghezza totale massima 4.*

Come mostrato da Pólya nella sua pubblicazione [2], il Teorema 2 è, a sua volta, una conseguenza di un celebre risultato, dovuto a Chebyshev. Per rendere questo capitolo indipendente, abbiamo incluso una dimostrazione nell'appendice (seguendo la splendida formulazione di Pólya e Szegő).

Pavnuty Chebyshev su un francobollo sovietico del 1946

Teorema di Chebyshev.
Sia $p(x)$ un polinomio reale di grado $n \geq 1$ con coefficiente direttivo 1. Allora

$$\max_{-1\leq x\leq 1} |p(x)| \ \geq \ \frac{1}{2^{n-1}}.$$

Notiamo dapprima la seguente conclusione immediata.

Corollario. *Sia $p(x)$ un polinomio reale di grado $n \geq 1$ con coefficiente direttivo 1, e si supponga che $|p(x)| \leq 2$ per ogni x nell'intervallo $[a, b]$. Allora $b - a \leq 4$.*

■ **Dimostrazione.** Si consideri la sostituzione $y = \frac{2}{b-a}(x - a) - 1$. Essa trasforma l'intervallo $[a, b]$ delle x nell'intervallo $[-1, 1]$ delle y. Il polinomio corrispondente

$$q(y) \ = \ p(\tfrac{b-a}{2}(y + 1) + a)$$

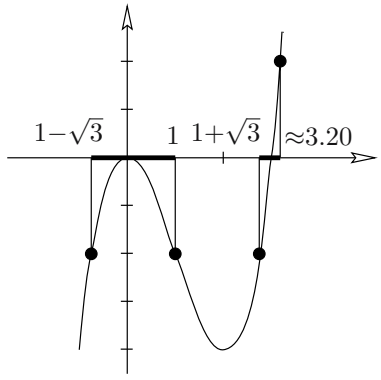
ha coefficiente direttivo $(\frac{b-a}{2})^n$ e soddisfa

$$\max_{-1\leq y\leq 1} |q(y)| \ = \ \max_{a\leq x\leq b} |p(x)|.$$

Mediante il teorema di Chebyshev deduciamo

$$2 \;\geq\; \max_{a\leq x\leq b} |p(x)| \;\geq\; (\tfrac{b-a}{2})^n \tfrac{1}{2^{n-1}} \;=\; 2(\tfrac{b-a}{4})^n$$

e pertanto $b-a \leq 4$, come desiderato. □



Per il polinomio $p(x) = x^2(x-3)$ otteniamo $\mathcal{P} = [1-\sqrt{3}, 1] \cup [1+\sqrt{3}, \approx 3.2]$

Questo corollario ci porta già molto vicini all'affermazione del Teorema 2. Se l'insieme $\mathcal{P} = \{x : |p(x)| \leq 2\}$ è un *intervallo*, allora la lunghezza di $\mathcal{P}$ è al massimo 4. L'insieme $\mathcal{P}$ tuttavia può non essere un intervallo, come nell'esempio riportato a margine, dove $\mathcal{P}$ consiste in due intervalli.

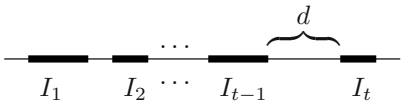
Cosa possiamo dire di $\mathcal{P}$? Dal momento che $p(x)$ è una funzione continua, sappiamo in ogni caso che $\mathcal{P}$ è l'unione di intervalli chiusi disgiunti $I_1, I_2, \ldots$ e che $p(x)$ assume il valore 2 o -2 ad ogni estremo di un intervallo I_j. Ciò implica che vi è solo un numero finito di intervalli $I_1, \ldots, I_t$, poiché $p(x)$ può assumere un valore dato solo un numero finito di volte.

L'idea magnifica di Pólya fu di costruire un altro polinomio $\tilde{p}(x)$ di grado n, di nuovo con coefficiente direttivo 1, tale che $\widetilde{\mathcal{P}} = \{x : |\tilde{p}(x)| \leq 2\}$ sia un *intervallo* di lunghezza non inferiore a $\ell(I_1) + \ldots + \ell(I_t)$. Il corollario dimostra allora che $\ell(I_1) + \ldots + \ell(I_t) \leq \ell(\widetilde{\mathcal{P}}) \leq 4$; ed abbiamo terminato.

■ **Dimostrazione del Teorema 2.** Si consideri $p(x) = (x-a_1)\cdots(x-a_n)$ con $\mathcal{P} = \{x \in \mathbb{R} : |p(x)| \leq 2\} = I_1 \cup \ldots \cup I_t$, dove disponiamo gli intervalli I_j in modo tale che I_1 sia il più a sinistra e I_t il più a destra. Dapprima affermiamo che ogni intervallo I_j contiene una radice di $p(x)$. Sappiamo che $p(x)$ assume i valori 2 o -2 agli estremi di I_j. Se un valore è 2 e l'altro -2, allora vi è certamente una radice in I_j. Pertanto supponiamo $p(x) = 2$ in entrambi gli estremi (il caso -2 essendo analogo). Sia $b \in I_j$ un punto in cui $p(x)$ assume il proprio minimo in I_j. Allora $p'(b) = 0$ e $p''(b) \geq 0$. Se $p''(b) = 0$, allora b è una radice multipla di $p'(x)$, e pertanto una radice di $p(x)$ in base alla Proposizione 1 del riquadro alla pagina seguente. Se, d'altro canto, $p''(b) > 0$, allora deduciamo che $p(b) \leq 0$ dalla Proposizione 2 nello stesso riquadro. Pertanto o $p(b) = 0$ ed abbiamo la nostra radice, oppure $p(b) < 0$ ed otteniamo una radice nell'intervallo da b ad uno dei punti estremi di I_j.

Ecco l'idea conclusiva della dimostrazione. Siano $I_1, \ldots, I_t$ gli intervalli come prima, e si supponga che l'intervallo più a destra I_t contenga m radici di $p(x)$, contate con le loro molteplicità. Se $m = n$, allora I_t è l'unico intervallo (in base a quanto abbiamo appena dimostrato) ed abbiamo terminato. Si supponga pertanto che $m < n$ e sia d la distanza fra I_{t-1} e I_t come nella figura. Siano $b_1, \ldots, b_m$ le radici di $p(x)$ che giacciono in I_t e $c_1, \ldots c_{n-m}$ quelle rimanenti. Scriviamo ora $p(x) = q(x)r(x)$ dove $q(x) = (x-b_1)\cdots(x-b_m)$ e $r(x) = (x-c_1)\cdots(x-c_{n-m})$ e poniamo $p_1(x) = q(x+d)r(x)$. Il polinomio $p_1(x)$ è nuovamente di grado n, monico. Per $x \in I_1 \cup \ldots \cup I_{t-1}$ abbiamo $|x+d-b_i| < |x-b_i|$ per ogni i e pertanto $|q(x+d)| < |q(x)|$. Ne segue che

$$|p_1(x)| \;\leq\; |p(x)| \;\leq\; 2 \qquad \text{per } x \in I_1 \cup \ldots \cup I_{t-1}.$$



Se, d'altro canto, $x \in I_t$, allora troviamo $|r(x-d)| \leq |r(x)|$ e quindi

$$|p_1(x-d)| \;=\; |q(x)||r(x-d)| \;\leq\; |p(x)| \;\leq\; 2,$$

il che significa che $I_t - d \subseteq \mathcal{P}_1 = \{x : |p_1(x)| \leq 2\}$.
Riassumendo, vediamo che $\mathcal{P}_1$ contiene $I_1 \cup \ldots \cup I_{t-1} \cup (I_t - d)$ e pertanto ha una lunghezza totale almeno grande quanto $\mathcal{P}$. Si noti ora che col passaggio da $p(x)$ a $p_1(x)$ gli intervalli I_{t-1} e $I_t - d$ si fondono in un unico intervallo. Concludiamo che gli intervalli $J_1, \ldots, J_s$ di $p_1(x)$ che compongono $\mathcal{P}_1$ hanno una lunghezza totale minima di $\ell(I_1) + \ldots + \ell(I_t)$ e che l'intervallo più a destra J_s contiene più di m radici di $p_1(x)$. Ripetendo questo procedimento al massimo $t-1$ volte, arriviamo infine ad un polinomio $\tilde{p}(x)$ in corrispondenza del quale $\widetilde{\mathcal{P}} = \{x : |\tilde{p}(x)| \leq 2\}$ è un intervallo di lunghezza $\ell(\widetilde{\mathcal{P}}) \geq \ell(I_1) + \ldots \ell(I_t)$ e la dimostrazione è completa. □

Due risultati sui polinomi con radici reali

Sia $p(x)$ un polinomio non costante con sole radici reali.

Proposizione 1. *Se b è una radice multipla di $p'(x)$, allora b è anche una radice di $p(x)$.*

■ **Dimostrazione.** Siano $b_1, \ldots, b_r$ le radici di $p(x)$ con molteplicità $s_1, \ldots, s_r$, $\sum_{j=1}^{r} s_j = n$. Dalla relazione $p(x) = (x - b_j)^{s_j} h(x)$ deduciamo che b_j è una radice di $p'(x)$ se $s_j \geq 2$ e la molteplicità di b_j in $p'(x)$ è $s_j - 1$. Inoltre, vi è una radice di $p'(x)$ tra b_1 e b_2, un'altra radice tra b_2 e $b_3, \ldots$, ed una tra b_{r-1} e b_r e tutte queste radici devono essere radici *semplici*, poiché $\sum_{j=1}^{r}(s_j - 1) + (r-1)$ dà già il grado $n-1$ di $p'(x)$. Di conseguenza, le radici *multiple* di $p'(x)$ possono apparire soltanto fra le radici di $p(x)$. □

Proposizione 2. *Abbiamo che $p'(x)^2 \geq p(x)p''(x)$ per ogni $x \in \mathbb{R}$.*

■ **Dimostrazione.** Se $x = a_i$ è una radice di $p(x)$, allora non c'è nulla da dimostrare. Si supponga quindi che x non sia una radice. La regola del prodotto del calcolo differenziale dà

$$p'(x) \;=\; \sum_{k=1}^{n} \frac{p(x)}{x - a_k}, \quad \text{ovvero,} \quad \frac{p'(x)}{p(x)} \;=\; \sum_{k=1}^{n} \frac{1}{x - a_k}.$$

Derivando ancora abbiamo

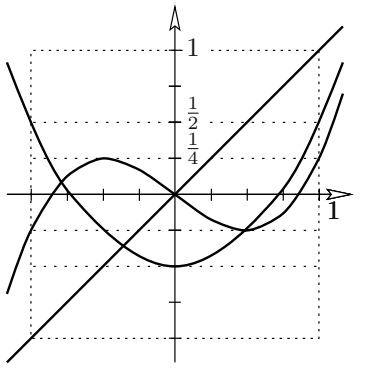
$$\frac{p''(x)p(x) - p'(x)^2}{p(x)^2} \;=\; -\sum_{k=1}^{n} \frac{1}{(x-a_k)^2} \;<\; 0. \qquad \square$$

Appendice: il teorema di Chebyshev

Teorema. *Sia $p(x)$ un polinomio reale di grado $n \geq 1$, monico. Allora*

$$\max_{-1\leq x\leq 1} |p(x)| \;\geq\; \frac{1}{2^{n-1}}.$$

Prima di iniziare, guardiamo qualche esempio dove abbiamo l'uguaglianza. A margine sono raffigurati i grafici di polinomi di grado 1, 2 e 3, dove in ogni caso abbiamo uguaglianza. In effetti, vedremo che per ogni grado vi è esattamente un polinomio con uguaglianza nel teorema di Chebyshev.



I polinomi $p_1(x) = x$, $p_2(x) = x^2 - \frac{1}{2}$ e $p_3(x) = x^3 - \frac{3}{4}x$ danno luogo ad un'uguaglianza nel teorema di Chebyshev

■ **Dimostrazione.** Si consideri un polinomio reale $p(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$, monico. Poiché siamo interessati all'intervallo $-1 \leq x \leq 1$, poniamo $x = \cos\vartheta$ ed indichiamo con $g(\vartheta) := p(\cos\vartheta)$ il polinomio risultante in $\cos\vartheta$,

$$g(\vartheta) = (\cos\vartheta)^n + a_{n-1}(\cos\vartheta)^{n-1} + \ldots + a_0. \tag{1}$$

La dimostrazione procede ora secondo le due fasi seguenti, ognuna delle quali costituisce un risultato classico ed interessante di per sé.

(A) Esprimiamo $g(\vartheta)$ come un cosiddetto *polinomio trigonometrico nel coseno*, ovvero, un polinomio della forma

$$g(\vartheta) \;=\; b_n \cos n\vartheta + b_{n-1}\cos(n-1)\vartheta + \ldots + b_1\cos\vartheta + b_0 \tag{2}$$

con $b_k \in \mathbb{R}$, e dimostriamo che il suo coefficiente direttivo è $b_n = \frac{1}{2^{n-1}}$.

(B) Dato un polinomio nel coseno $h(\vartheta)$ di ordine n (che significa che λ_n è il più alto coefficiente non nullo)

$$h(\vartheta) \;=\; \lambda_n \cos n\vartheta + \lambda_{n-1}\cos(n-1)\vartheta + \ldots + \lambda_0, \tag{3}$$

mostriamo che $|\lambda_n| \;\leq\; \max|h(\vartheta)|$; che applicato a $g(\vartheta)$ dimostrerà il teorema.

Dimostrazione di (A). Per passare da (1) alla rappresentazione (2), dobbiamo esprimere tutte le potenze $(\cos\vartheta)^k$ come polinomi nel coseno. Ad esempio, il teorema della somma per il coseno dà

$$\cos 2\vartheta \;=\; \cos^2\vartheta - \sin^2\vartheta \;=\; 2\cos^2\vartheta - 1,$$

cosicché $\cos^2\vartheta = \frac{1}{2}\cos 2\vartheta + \frac{1}{2}$. Nel caso di una potenza arbitraria $(\cos\vartheta)^k$, utilizziamo i numeri complessi, tramite la relazione $e^{ix} = \cos x + i\sin x$. Gli e^{ix} sono i numeri complessi di valore assoluto 1 (si veda il riquadro sulle radici complesse dell'unità a pagina 29). In particolare, questo dà

$$e^{in\vartheta} \;=\; \cos n\vartheta + i\sin n\vartheta. \tag{4}$$

D'altro canto,

$$e^{in\vartheta} \;=\; (e^{i\vartheta})^n \;=\; (\cos\vartheta + i\sin\vartheta)^n. \tag{5}$$

Uguagliando le parti reali in (4) e (5) ed usando le relazioni $i^{4\ell+2} = -1$, $i^{4\ell} = 1$ e $\sin^2\theta = 1 - \cos^2\theta$, otteniamo

$$\begin{aligned}\cos n\vartheta &= \sum_{\ell\geq 0}\binom{n}{4\ell}(\cos\vartheta)^{n-4\ell}(1-\cos^2\vartheta)^{2\ell} \\ &\quad -\sum_{\ell\geq 0}\binom{n}{4\ell+2}(\cos\vartheta)^{n-4\ell-2}(1-\cos^2\vartheta)^{2\ell+1}.\end{aligned} \tag{6}$$

Concludiamo che $\cos n\vartheta$ è un polinomio in $\cos\vartheta$,

$$\cos n\vartheta = c_n(\cos\vartheta)^n + c_{n-1}(\cos\vartheta)^{n-1} + \ldots + c_0. \tag{7}$$

$\sum_{k\geq 0}\binom{n}{2k} = 2^{n-1}$ per $n > 0$: ogni sottoinsieme di $\{1,2,\ldots,n-1\}$ dà un insieme di dimensione *pari* di $\{1,2,\ldots,n\}$ se aggiungiamo l'elemento n "quando necessario"

Da (6) otteniamo per il coefficiente più elevato

$$c_n = \sum_{\ell\geq 0}\binom{n}{4\ell} + \sum_{\ell\geq 0}\binom{n}{4\ell+2} = 2^{n-1}.$$

Ora invertiamo il nostro argomento. Supponendo per induzione che per $k < n$, $(\cos\vartheta)^k$ possa esprimersi come un polinomio nel coseno di ordine k, deduciamo da (7) che $(\cos\vartheta)^n$ si può scrivere come un polinomio nel coseno di ordine n con coefficiente direttivo $b_n = \frac{1}{2^{n-1}}$.

Dimostrazione di (B). Sia $h(\vartheta)$ un polinomio nel coseno di ordine n come in (3), e si supponga senza perdere generalità che $\lambda_n > 0$. Ora poniamo $m(\vartheta) := \lambda_n \cos n\vartheta$ e troviamo

$$m(\tfrac{k}{n}\pi) = (-1)^k\lambda_n \qquad \text{per } k = 0,1,\ldots,n.$$

Procedendo per contraddizione, supponiamo che $\max|h(\vartheta)| < \lambda_n$. Allora

$$m(\tfrac{k}{n}\pi) - h(\tfrac{k}{n}\pi) = (-1)^k\lambda_n - h(\tfrac{k}{n}\pi)$$

è positivo per k pari e negativo per k dispari nell'intervallo $0 \leq k \leq n$. Concludiamo che $m(\vartheta) - h(\vartheta)$ ha almeno n radici nell'intervallo $[0,\pi]$. Ma ciò non può essere poiché $m(\vartheta) - h(\vartheta)$ è un polinomio nel coseno di ordine $n-1$, che può essere scritto nella forma (1) e pertanto ha al massimo $n-1$ radici.

La dimostrazione di **(B)** e dunque del teorema di Chebyshev è completa. □

Il lettore può ora completare facilmente l'analisi, dimostrando che $g_n(\vartheta) := \frac{1}{2^{n-1}}\cos n\vartheta$ è l'*unico* polinomio nel coseno di ordine n, monico, che soddisfa l'uguaglianza $\max|g(\vartheta)| = \frac{1}{2^{n-1}}$.

I polinomi $T_n(x) = \cos n\vartheta$, $x = \cos\vartheta$ sono chiamati *polinomi di Chebyshev* (di prima specie); dunque $\frac{1}{2^{n-1}}T_n(x)$ è l'unico polinomio monico di grado n per cui valga l'uguaglianza nel teorema di Chebyshev.

Bibliografia

[1] P. L. CEBYCEV: *Œuvres,* Vol. I, Acad. Imperiale des Sciences, St. Petersburg 1899, pp. 387-469.

[2] G. PÓLYA: *Beitrag zur Verallgemeinerung des Verzerrungssatzes auf mehrfach zusammenhängenden Gebieten,* Sitzungsber. Preuss. Akad. Wiss. Berlin (1928), 228-232; Collected Papers Vol. I, MIT Press 1974, 347-351.

[3] G. PÓLYA & G. SZEGŐ: *Problems and Theorems in Analysis, Vol. II,* Springer-Verlag, Berlin Heidelberg New York 1976; Reprint 1998.

Su un lemma di Littlewood e Offord

Capitolo 19

Nel loro lavoro sulla distribuzione delle radici di equazioni algebriche, nel 1943 Littlewood e Offord dimostrarono il seguente risultato:

> *Siano $a_1, a_2, \ldots, a_n$ numeri complessi tali che $|a_i| \geq 1$ per ogni i, e si considerino le 2^n combinazioni lineari $\sum_{i=1}^n \varepsilon_i a_i$ con $\varepsilon_i \in \{1, -1\}$. Allora il numero delle somme $\sum_{i=1}^n \varepsilon_i a_i$ che giacciono all'interno di un cerchio di raggio 1 non è più grande di*
>
> $$c\,\frac{2^n}{\sqrt{n}}\,\log n \quad \textit{per una costante } c > 0.$$

Qualche anno più tardi Paul Erdős migliorò questo limite eliminando il termine $\log n$ e, cosa più interessante, dimostrò che questo è, in effetti, una semplice conseguenza del teorema di Sperner (si veda a pagina 169).

Per avere un'idea di quest'argomentazione, consideriamo il caso in cui tutti gli a_i sono reali. Possiamo supporre che tutti gli a_i siano positivi (sostituendo a_i con $-a_i$ e ε_i con $-\varepsilon_i$ ogni qualvolta $a_i < 0$). Supponiamo ora che un insieme di combinazioni $\sum \varepsilon_i a_i$ giaccia all'interno di un intervallo di lunghezza 2. Sia $N = \{1, 2, \ldots, n\}$ l'insieme degli indici. Per ogni $\sum \varepsilon_i a_i$ poniamo $I := \{i \in N : \varepsilon_i = 1\}$. Ora se $I \subsetneqq I'$ per due di tali insiemi, allora concludiamo che

$$\sum \varepsilon_i' a_i - \sum \varepsilon_i a_i \;=\; 2 \sum_{i \in I' \setminus I} a_i \;\geq\; 2,$$

il che è una contraddizione. Pertanto gli insiemi I formano un'anti-catena, e concludiamo in base al teorema di Sperner che vi sono al massimo $\binom{n}{\lfloor n/2 \rfloor}$ di tali combinazioni. In base alla formula di Stirling (si veda a pagina 11) abbiamo

$$\binom{n}{\lfloor n/2 \rfloor} \;\leq\; c\,\frac{2^n}{\sqrt{n}} \quad \text{per un opportuno } c > 0.$$

Per n pari e per ogni $a_i = 1$ otteniamo $\binom{n}{n/2}$ combinazioni $\sum_{i=1}^n \varepsilon_i a_i$ la cui somma valga 0. Guardando all'intervallo $(-1, 1)$ troviamo pertanto che il numero binomiale fornisce la maggiorazione *esatta*.

Nella stessa pubblicazione Erdős congetturò che $\binom{n}{\lfloor n/2 \rfloor}$ fosse la giusta maggiorazione anche per i numeri complessi (egli potè solo dimostrare $c\,2^n n^{-1/2}$ per un opportuno c) e che in effetti la stessa maggiorazione è valida per vettori $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ con $|\boldsymbol{a}_i| \geq 1$ in uno spazio di Hilbert reale, quando il cerchio di raggio 1 è sostituito con una palla aperta di raggio 1.

John E. Littlewood

Il teorema di Sperner. *Ogni anti-catena di sottoinsiemi di un insieme di cardinalità n ha dimensione massima $\binom{n}{\lfloor n/2 \rfloor}$*

Erdős aveva ragione, ma ci vollero vent'anni prima che Gyula Katona e Daniel Kleitman giungessero indipendentemente ad una dimostrazione per i numeri complessi (o, il che è lo stesso, per il piano $\mathbb{R}^2$). Le loro dimostrazioni facevano uso esplicito della bidimensionalità del piano e non era del tutto chiaro come potessero essere estese per coprire spazi vettoriali reali di dimensione finita.

Tuttavia nel 1970 Kleitman dimostrò la congettura completa in spazi di Hilbert con un argomento di semplicità sbalorditiva. In effetti, egli dimostrò anche più di questo. Il suo argomento è un esempio lampante di quello che si può fare quando si trova la giusta ipotesi d'induzione.

Una parola di conforto per tutti i lettori che non conoscono gli spazi di Hilbert: non ne abbiamo realmente bisogno. Dal momento che trattiamo soltanto un numero finito di vettori $\boldsymbol{a}_i$, basta considerare lo spazio reale $\mathbb{R}^d$ con il solito prodotto scalare. Ecco il risultato di Kleitman.

Teorema. *Siano $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ vettori in $\mathbb{R}^d$, ognuno di lunghezza minima 1, e siano $R_1, \ldots, R_k$ k regioni aperte di $\mathbb{R}^d$, dove $|\boldsymbol{x} - \boldsymbol{y}| < 2$ per ogni $\boldsymbol{x}, \boldsymbol{y}$ che giacciano nella stessa regione R_i. Allora il numero di combinazioni lineari $\sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i$, $\varepsilon_i \in \{1, -1\}$, che possono trovarsi nell'unione $\bigcup_i R_i$ di tali regioni è al più uguale alla somma dei k più grandi coefficienti binomiali $\binom{n}{j}$.*
In particolare, per $k = 1$ abbiamo la maggiorazione $\binom{n}{\lfloor n/2 \rfloor}$.

Prima di dimostrarlo si noti che la maggiorazione è esatta se

$$\boldsymbol{a}_1 \;=\; \ldots \;=\; \boldsymbol{a}_n \;=\; \boldsymbol{a} \;=\; (1, 0, \ldots, 0)^T.$$

In effetti, per n pari otteniamo $\binom{n}{n/2}$ somme uguali a 0, $\binom{n}{n/2-1}$ somme uguali a $(-2)\boldsymbol{a}$, $\binom{n}{n/2+1}$ somme uguali a $2\boldsymbol{a}$, e così via. Scegliendo palle di raggio 1 di centro

$$-2\lceil \tfrac{k-1}{2} \rceil \boldsymbol{a}, \quad \ldots \quad (-2)\boldsymbol{a}, \quad \mathbf{0}, \quad 2\boldsymbol{a}, \quad \ldots \quad 2\lfloor \tfrac{k-1}{2} \rfloor \boldsymbol{a},$$

otteniamo

$$\binom{n}{\lfloor \frac{n-k+1}{2} \rfloor} + \ldots + \binom{n}{\frac{n-2}{2}} + \binom{n}{\frac{n}{2}} + \binom{n}{\frac{n+2}{2}} + \ldots + \binom{n}{\lfloor \frac{n+k-1}{2} \rfloor}$$

somme che giacciono in queste k palle; e questa è l'espressione promessa, poiché i coefficienti binomiali più grandi sono centrati nel mezzo (si veda a pagina 12). Un ragionamento analogo vale nel caso in cui n sia dispari.

■ **Dimostrazione.** Da questo momento in poi possiamo supporre, senza perdita di generalità, che le regioni R_i siano disgiunte. La chiave della dimostrazione è la relazione ricorsiva fra i coefficienti binomiali, che ci dice qual è la relazione tra i più grandi coefficienti binomiali di n e $n-1$. Si ponga $r = \lfloor \frac{n-k+1}{2} \rfloor, s = \lfloor \frac{n+k-1}{2} \rfloor$, allora $\binom{n}{r}, \binom{n}{r+1}, \ldots, \binom{n}{s}$ sono i k coefficienti binomiali più grandi per n. La relazione ricorsiva $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$ implica

Erdős aveva ragione, ma ci vollero vent'anni prima che Gyula Katona e Daniel Kleitman giungessero indipendentemente ad una dimostrazione per i numeri complessi (o, il che è lo stesso, per il piano $\mathbb{R}^2$). Le loro dimostrazioni facevano uso esplicito della bidimensionalità del piano e non era del tutto chiaro come potessero essere estese per coprire spazi vettoriali reali di dimensione finita.

Tuttavia nel 1970 Kleitman dimostrò la congettura completa in spazi di Hilbert con un argomento di semplicità sbalorditiva. In effetti, egli dimostrò anche più di questo. Il suo argomento è un esempio lampante di quello che si può fare quando si trova la giusta ipotesi d'induzione.

Una parola di conforto per tutti i lettori che non conoscono gli spazi di Hilbert: non ne abbiamo realmente bisogno. Dal momento che trattiamo soltanto un numero finito di vettori $\boldsymbol{a}_i$, basta considerare lo spazio reale $\mathbb{R}^d$ con il solito prodotto scalare. Ecco il risultato di Kleitman.

Teorema. *Siano $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ vettori in $\mathbb{R}^d$, ognuno di lunghezza minima 1, e siano $R_1, \ldots, R_k$ k regioni aperte di $\mathbb{R}^d$, dove $|\boldsymbol{x} - \boldsymbol{y}| < 2$ per ogni $\boldsymbol{x}, \boldsymbol{y}$ che giacciano nella stessa regione R_i. Allora il numero di combinazioni lineari $\sum_{i=1}^n \varepsilon_i \boldsymbol{a}_i$, $\varepsilon_i \in \{1, -1\}$, che possono trovarsi nell'unione $\bigcup_i R_i$ di tali regioni è al più uguale alla somma dei k più grandi coefficienti binomiali $\binom{n}{j}$.*
In particolare, per $k = 1$ abbiamo la maggiorazione $\binom{n}{\lfloor n/2 \rfloor}$.

Prima di dimostrarlo si noti che la maggiorazione è esatta se

$$\boldsymbol{a}_1 \;=\; \ldots \;=\; \boldsymbol{a}_n \;=\; \boldsymbol{a} \;=\; (1, 0, \ldots, 0)^T.$$

In effetti, per n pari otteniamo $\binom{n}{n/2}$ somme uguali a 0, $\binom{n}{n/2-1}$ somme uguali a $(-2)\boldsymbol{a}$, $\binom{n}{n/2+1}$ somme uguali a $2\boldsymbol{a}$, e così via. Scegliendo palle di raggio 1 di centro

$$-2\lceil \tfrac{k-1}{2} \rceil \boldsymbol{a}, \quad \ldots \quad (-2)\boldsymbol{a}, \quad \mathbf{0}, \quad 2\boldsymbol{a}, \quad \ldots \quad 2\lfloor \tfrac{k-1}{2} \rfloor \boldsymbol{a},$$

otteniamo

$$\binom{n}{\lfloor \frac{n-k+1}{2} \rfloor} + \ldots + \binom{n}{\frac{n-2}{2}} + \binom{n}{\frac{n}{2}} + \binom{n}{\frac{n+2}{2}} + \ldots + \binom{n}{\lfloor \frac{n+k-1}{2} \rfloor}$$

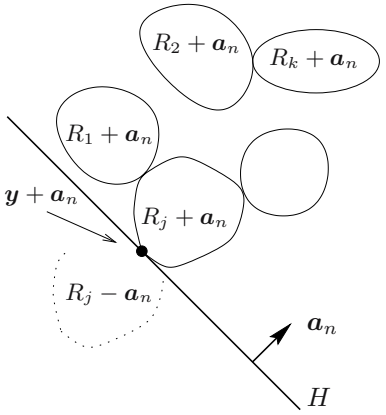
somme che giacciono in queste k palle; e questa è l'espressione promessa, poiché i coefficienti binomiali più grandi sono centrati nel mezzo (si veda a pagina 12). Un ragionamento analogo vale nel caso in cui n sia dispari.

■ **Dimostrazione.** Da questo momento in poi possiamo supporre, senza perdita di generalità, che le regioni R_i siano disgiunte. La chiave della dimostrazione è la relazione ricorsiva fra i coefficienti binomiali, che ci dice qual è la relazione tra i più grandi coefficienti binomiali di n e $n-1$. Si ponga $r = \lfloor \frac{n-k+1}{2} \rfloor, s = \lfloor \frac{n+k-1}{2} \rfloor$, allora $\binom{n}{r}, \binom{n}{r+1}, \ldots, \binom{n}{s}$ sono i k coefficienti binomiali più grandi per n. La relazione ricorsiva $\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}$ implica

$$\begin{aligned}\sum_{i=r}^{s}\binom{n}{i} &= \sum_{i=r}^{s}\binom{n-1}{i}+\sum_{i=r}^{s}\binom{n-1}{i-1}\\ &= \sum_{i=r}^{s}\binom{n-1}{i}+\sum_{i=r-1}^{s-1}\binom{n-1}{i} \qquad (1)\\ &= \sum_{i=r-1}^{s}\binom{n-1}{i}+\sum_{i=r}^{s-1}\binom{n-1}{i},\end{aligned}$$

ed un semplice calcolo mostra che la prima somma addiziona i $k+1$ più grandi coefficienti binomiali $\binom{n-1}{i}$, la seconda i $k-1$ più grandi.

La dimostrazione di Kleitman procede per induzione su n, il caso $n=1$ essendo banale. Alla luce di (1) dobbiamo soltanto dimostrare per il passo di induzione che le combinazioni lineari di $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n$ che giacciono in k regioni disgiunte possono essere trasformate *biiettivamente* in combinazioni di $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{n-1}$ che giacciono in $k+1$ o $k-1$ regioni.



Proposizione. *Almeno una delle regioni traslate $R_j-\boldsymbol{a}_n$ è disgiunta da tutte le regioni traslate $R_1+\boldsymbol{a}_n,\ldots,R_k+\boldsymbol{a}_n$.*

Per dimostrare questo risultato, si consideri l'iperpiano $H=\{\boldsymbol{x}:\langle\boldsymbol{a}_n,\boldsymbol{x}\rangle = c\}$ ortogonale a $\boldsymbol{a}_n$, che contiene tutte le traslate $R_i+\boldsymbol{a}_n$ dalla parte caratterizzata da $\langle\boldsymbol{a}_n,\boldsymbol{x}\rangle\geq c$, e che tocca la chiusura di una regione, diciamo $R_j+\boldsymbol{a}_n$. Tale iperpiano esiste poiché le regioni sono limitate. Ora per ogni $\boldsymbol{x}\in R_j$ e $\boldsymbol{y}$ nella chiusura di R_j, $|\boldsymbol{x}-\boldsymbol{y}|<2$ poiché R_j è aperta. Vogliamo dimostrare che $R_j-\boldsymbol{a}_n$ giace dall'altra parte di H. Supponiamo, al contrario, che $\langle\boldsymbol{a}_n,\boldsymbol{x}-\boldsymbol{a}_n\rangle\geq c$ per qualche $\boldsymbol{x}\in R_j$ data, ovvero, $\langle\boldsymbol{a}_n,\boldsymbol{x}\rangle\geq|\boldsymbol{a}_n|^2+c$.

Sia $\boldsymbol{y}+\boldsymbol{a}_n$ un punto in cui H tocca $R_j+\boldsymbol{a}_n$, allora $\boldsymbol{y}$ è nella chiusura di R_j, e $\langle\boldsymbol{a}_n,\boldsymbol{y}+\boldsymbol{a}_n\rangle=c$, ovvero, $\langle\boldsymbol{a}_n,-\boldsymbol{y}\rangle=|\boldsymbol{a}_n|^2-c$. Pertanto

$$\langle\boldsymbol{a}_n,\boldsymbol{x}-\boldsymbol{y}\rangle \;\geq\; 2|\boldsymbol{a}_n|^2,$$

ed inferiamo, grazie alla disuguaglianza di Cauchy-Schwarz,

$$2|\boldsymbol{a}_n|^2 \;\leq\; \langle\boldsymbol{a}_n,\boldsymbol{x}-\boldsymbol{y}\rangle \;\leq\; |\boldsymbol{a}_n||\boldsymbol{x}-\boldsymbol{y}|,$$

e dunque (con $|\boldsymbol{a}_n|\geq 1$) otteniamo $2\leq 2|\boldsymbol{a}_n|\leq|\boldsymbol{x}-\boldsymbol{y}|$, il che è una contraddizione.

Il resto è facile. Classifichiamo le combinazioni $\sum\varepsilon_i\boldsymbol{a}_i$ che vengono a giacere in $R_1\cup\ldots\cup R_k$ come segue. Nella Classe 1 mettiamo tutte le $\sum_{i=1}^{n}\varepsilon_i\boldsymbol{a}_i$ con $\varepsilon_n=-1$ e tutte le $\sum_{i=1}^{n}\varepsilon_i\boldsymbol{a}_i$ con $\varepsilon_n=1$ giacenti in R_j, mentre nella Classe 2 buttiamo le restanti combinazioni $\sum_{i=1}^{n}\varepsilon_i\boldsymbol{a}_i$ con $\varepsilon_n=1$, non in R_j. Ne segue che le combinazioni $\sum_{i=1}^{n-1}\varepsilon_i\boldsymbol{a}_i$ corrispondenti alla Classe 1 giacciono nelle $k+1$ regioni disgiunte $R_1+\boldsymbol{a}_n,\ldots,R_k+\boldsymbol{a}_n$ e $R_j-\boldsymbol{a}_n$, mentre le combinazioni $\sum_{i=1}^{n-1}\varepsilon_i\boldsymbol{a}_i$ corrispondenti alla Classe 2 giacciono nelle $k-1$ regioni disgiunte $R_1-\boldsymbol{a}_n,\ldots,R_k-\boldsymbol{a}_n$ senza $R_j-\boldsymbol{a}_n$. Per induzione, la Classe 1 contiene al massimo $\sum_{i=r-1}^{s}\binom{n-1}{i}$

combinazioni, mentre la Classe 2 contiene al massimo $\sum_{i=r}^{s-1} \binom{n-1}{i}$ combinazioni — e grazie a (1) questa è la dimostrazione completa, direttamente dal *Libro*. □

Bibliografia

[1] P. ERDŐS: *On a lemma of Littlewood and Offord,* Bulletin Amer. Math. Soc. **51** (1945), 898-902.

[2] G. KATONA: *On a conjecture of Erdős and a stronger form of Sperner's theorem,* Studia Sci. Math. Hungar. **1** (1966), 59-63.

[3] D. KLEITMAN: *On a lemma of Littlewood and Offord on the distribution of certain sums,* Math. Zeitschrift **90** (1965), 251-259.

[4] D. KLEITMAN: *On a lemma of Littlewood and Offord on the distributions of linear combinations of vectors,* Advances Math. **5** (1970), 155-157.

[5] J. E. LITTLEWOOD & A. C. OFFORD: *On the number of real roots of a random algebraic equation III,* Mat. USSR Sb. **12** (1943), 277-285.

La funzione cotangente e il trucco di Herglotz

Capitolo 20

Fra le formule basate su funzioni elementari qual è la più interessante? In uno splendido articolo [2], di cui seguiamo da vicino l'esposizione, Jürgen Elstrodt mette al primo posto lo sviluppo in serie della funzione contangente:

$$\pi \cot \pi x \;=\; \frac{1}{x} + \sum_{n=1}^{\infty} \Big(\frac{1}{x+n} + \frac{1}{x-n} \Big) \qquad (x \in \mathbb{R} \backslash \mathbb{Z}).$$

Questa formula elegante fu dimostrata da Eulero nel §178 della sua *Introductio in Analysin Infinitorum* del 1748 e certamente si annovera fra i suoi migliori successi. Possiamo anche scriverla ancora più elegantemente come

$$\pi \cot \pi x \;=\; \lim_{N \to \infty} \sum_{n=-N}^{N} \frac{1}{x+n} \tag{1}$$

ma bisogna notare che la stima della somma $\sum_{n \in \mathbb{Z}} \frac{1}{x+n}$ è un po' pericolosa, poiché la serie non è assolutamente convergente, dunque il suo valore dipende da una scelta giudiziosa dell'ordine della sommatoria.

Deriveremo (1) mediante un argomento di semplicità sbalorditiva attribuito a Gustav Herglotz — il "trucco di Herglotz". Per cominciare, si ponga

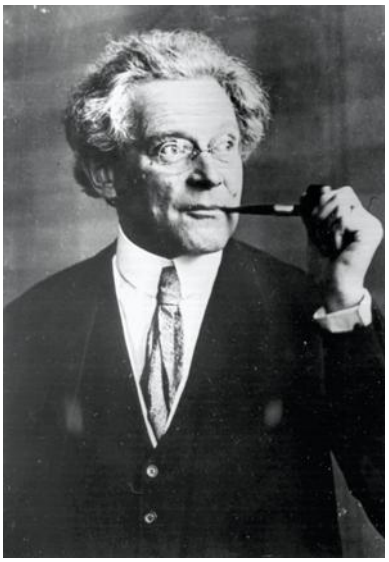
$$f(x) \;:=\; \pi \cot \pi x, \qquad g(x) \;:=\; \lim_{N \to \infty} \sum_{n=-N}^{N} \frac{1}{x+n},$$

e cerchiamo di derivare un numero abbastanza grande di proprietà comuni di queste due funzioni in modo da poter concludere che alla fine esse debbano coincidere...

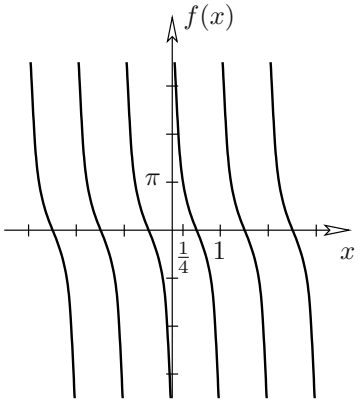
(A) Le funzioni f e g sono definite per tutti i valori non interi e sono continue in essi.

Per la funzione cotangente $f(x) = \pi \cot \pi x = \pi \frac{\cos \pi x}{\sin \pi x}$, ciò è evidente (si veda la figura). Per $g(x)$, usiamo dapprima l'identità $\frac{1}{x+n} + \frac{1}{x-n} = -\frac{2x}{n^2-x^2}$ per riscrivere la formula di Eulero come

$$\pi \cot \pi x \;=\; \frac{1}{x} - \sum_{n=1}^{\infty} \frac{2x}{n^2 - x^2}\,. \tag{2}$$



Gustav Herglotz



La funzione $f(x) = \pi \cot \pi x$

Pertanto per dimostrare **(A)** dobbiamo dimostrare che per ogni $x \notin \mathbb{Z}$ la serie

$$\sum_{n=1}^{\infty} \frac{1}{n^2 - x^2}$$

converge uniformemente in un intorno di x.

Per questo, non abbiamo problemi con il primo termine, per $n = 1$, né con i termini tali che $2n - 1 \leq x^2$, poiché ve ne è solo un numero finito. D'altro canto, per $n \geq 2$ e $2n - 1 > x^2$, ovvero $n^2 - x^2 > (n-1)^2 > 0$, gli addendi sono limitati da

$$0 \;<\; \frac{1}{n^2 - x^2} \;<\; \frac{1}{(n-1)^2},$$

e tale limite non vale solo per x stesso, ma anche per i valori in un intorno di x. Infine il fatto che $\sum \frac{1}{(n-1)^2}$ converge (a $\frac{\pi^2}{6}$, si veda a pagina 41) fornisce la convergenza uniforme necessaria alla dimostrazione di **(A)**.

(B) Sia f sia g sono *periodiche* di periodo 1, ovvero, $f(x+1) = f(x)$ e $g(x+1) = g(x)$ per ogni $x \in \mathbb{R}\backslash\mathbb{Z}$.

Poiché la cotangente ha periodo π, troviamo che f ha periodo 1 (si veda ancora la figura sopra). Per g argomentiamo come segue. Sia

$$g_N(x) \;:=\; \sum_{n=-N}^{N} \frac{1}{x+n},$$

allora

$$\begin{aligned} g_N(x+1) \;&=\; \sum_{n=-N}^{N} \frac{1}{x+1+n} \;=\; \sum_{n=-N+1}^{N+1} \frac{1}{x+n} \\ &=\; g_{N-1}(x) + \frac{1}{x+N} + \frac{1}{x+N+1}. \end{aligned}$$

Pertanto $g(x+1) = \lim_{N\to\infty} g_N(x+1) = \lim_{N\to\infty} g_{N-1}(x) = g(x)$.

(C) Sia f sia g sono funzioni *dispari*, ovvero abbiamo $f(-x) = -f(x)$ e $g(-x) = -g(x)$ per ogni $x \in \mathbb{R}\backslash\mathbb{Z}$.

La funzione f ha ovviamente questa proprietà e per g ci basta osservare che $g_N(-x) = -g_N(x)$.

I due fatti finali costituiscono il trucco di Herglotz: in primo luogo mostriamo che f e g soddisfano la stessa equazione funzionale e in secondo luogo che $h := f - g$ può essere estesa con continuità su tutto $\mathbb{R}$.

(D) Le due funzioni f e g soddisfano la stessa equazione funzionale: $f(\frac{x}{2}) + f(\frac{x+1}{2}) = 2f(x)$ e $g(\frac{x}{2}) + g(\frac{x+1}{2}) = 2g(x)$.

Per $f(x)$ ciò risulta dai teoremi della somma per le funzioni seno e coseno:

$$\begin{aligned} f(\tfrac{x}{2}) + f(\tfrac{x+1}{2}) &= \pi\left[\frac{\cos\frac{\pi x}{2}}{\sin\frac{\pi x}{2}} - \frac{\sin\frac{\pi x}{2}}{\cos\frac{\pi x}{2}}\right] \\ &= 2\pi\frac{\cos(\frac{\pi x}{2}+\frac{\pi x}{2})}{\sin(\frac{\pi x}{2}+\frac{\pi x}{2})} = 2\,f(x). \end{aligned}$$

Teoremi della somma:
$\sin(x+y) = \sin x\cos y + \cos x\sin y$
$\cos(x+y) = \cos x\cos y - \sin x\sin y$
$\Longrightarrow \sin(x+\frac{\pi}{2}) = \cos x$
$\cos(x+\frac{\pi}{2}) = -\sin x$
$\sin x = 2\sin\frac{x}{2}\cos\frac{x}{2}$
$\cos x = \cos^2\frac{x}{2} - \sin^2\frac{x}{2}$

L'equazione funzionale per g segue da

$$g_N(\tfrac{x}{2}) + g_N(\tfrac{x+1}{2}) = 2\,g_{2N}(x) + \frac{2}{x+2N+1},$$

che a sua volta deriva da

$$\frac{1}{\frac{x}{2}+n} + \frac{1}{\frac{x+1}{2}+n} = 2\Big(\frac{1}{x+2n} + \frac{1}{x+2n+1}\Big).$$

Ora consideriamo

$$h(x) = f(x) - g(x) = \pi\cot\pi x - \Big(\frac{1}{x} - \sum_{n=1}^{\infty}\frac{2x}{n^2-x^2}\Big). \qquad (3)$$

Sappiamo che h è una funzione continua su $\mathbb{R}\backslash\mathbb{Z}$ che soddisfa le proprietà **(B)**, **(C)**, **(D)**. Cosa succede nei valori interi? Dagli sviluppi in serie di seni e coseni, oppure applicando due volte la regola di de l'Hospital, troviamo

$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} \pm \ldots$
$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \pm \ldots$

$$\lim_{x\to 0}\Big(\cot x - \frac{1}{x}\Big) = \lim_{x\to 0}\frac{x\cos x - \sin x}{x\sin x} = 0,$$

e pertanto anche

$$\lim_{x\to 0}\Big(\pi\cot\pi x - \frac{1}{x}\Big) = 0.$$

Ma poiché l'ultima somma $\sum_{n=1}^{\infty}\frac{2x}{n^2-x^2}$ in (3) converge a 0 se $x \longrightarrow 0$, abbiamo in effetti $\lim_{x\to 0} h(x) = 0$, e pertanto per periodicità

$$\lim_{x\to n} h(x) = 0 \qquad \text{per ogni } n \in \mathbb{Z}.$$

Riassumendo, abbiamo dimostrato quanto segue:

(E) Ponendo $h(x) := 0$ per $x \in \mathbb{Z}$, h diventa una funzione continua su tutto $\mathbb{R}$ che gode delle proprietà enunciate in **(B)**, **(C)** e **(D)**.

Siamo pronti per il *coup de grâce* [NdT: in francese nel testo inglese]. Poiché h è una funzione continua periodica, essa ammette un massimo m. Sia x_0 un punto in $[0,1]$ tale che $h(x_0) = m$. Segue da **(D)** che

$$h(\tfrac{x_0}{2}) + h(\tfrac{x_0+1}{2}) = 2m,$$

e pertanto che $h(\frac{x_0}{2}) = m$. Iterando si ottiene $h(\frac{x_0}{2^n}) = m$ per ogni n, e pertanto $h(0) = m$ per continuità. Ma $h(0) = 0$ e quindi $m = 0$, ovvero,

$h(x) \leq 0$ per ogni $x \in \mathbb{R}$. Poiché $h(x)$ è una funzione *dispari*, $h(x) < 0$ è impossibile, dunque $h(x) = 0$ per ogni $x \in \mathbb{R}$ ed il teorema di Eulero è dimostrato. □

Un gran numero di corollari possono essere derivati da (1), il più celebre dei quali riguarda i valori della funzione zeta di Riemann per interi pari positivi (si veda l'Appendice al Capitolo 6),

$$\zeta(2k) \;=\; \sum_{n=1}^{\infty} \frac{1}{n^{2k}} \qquad (k \in \mathbb{N}). \tag{4}$$

Ora, per finire la nostra storia, vediamo come Eulero — qualche anno più tardi, nel 1755 — trattò la serie (4). Iniziamo dalla formula (2). Moltiplicando (2) per x e ponendo $y = \pi x$ troviamo per $|y| < \pi$:

$$\begin{aligned} y\cot y &= 1 - 2\sum_{n=1}^{\infty} \frac{y^2}{\pi^2 n^2 - y^2} \\ &= 1 - 2\sum_{n=1}^{\infty} \frac{y^2}{\pi^2 n^2} \, \frac{1}{1 - \left(\frac{y}{\pi n}\right)^2}. \end{aligned}$$

L'ultimo fattore è la somma di una serie geometrica, pertanto

$$\begin{aligned} y\cot y &= 1 - 2\sum_{n=1}^{\infty}\sum_{k=1}^{\infty} \Big(\frac{y}{\pi n}\Big)^{2k} \\ &= 1 - 2\sum_{k=1}^{\infty} \Big(\frac{1}{\pi^{2k}} \sum_{n=1}^{\infty} \frac{1}{n^{2k}}\Big) y^{2k}, \end{aligned}$$

ed abbiamo dimostrato il risultato notevole:

Per ogni $k \in \mathbb{N}$, il coefficiente di y^{2k} *nello sviluppo in serie di potenze di $y\cot y$ è uguale a*

$$\big[y^{2k}\big]\, y\cot y \;=\; -\frac{2}{\pi^{2k}} \sum_{n=1}^{\infty} \frac{1}{n^{2k}} \;=\; -\frac{2}{\pi^{2k}} \zeta(2k). \tag{5}$$

Esiste un altro modo, forse molto più canonico, di ottenere uno sviluppo in serie di $y\cot y$. Sappiamo dall'analisi che $e^{iy} = \cos y + i \sin y$, e pertanto

$$\cos y \;=\; \frac{e^{iy} + e^{-iy}}{2}, \qquad \sin y \;=\; \frac{e^{iy} - e^{-iy}}{2i},$$

il che dà

$$y\cot y \;=\; iy \frac{e^{iy} + e^{-iy}}{e^{iy} - e^{-iy}} \;=\; iy \frac{e^{2iy} + 1}{e^{2iy} - 1}\,.$$

Sostituiamo ora $z = 2iy$ ed otteniamo

$$y\cot y \;=\; \frac{z}{2}\, \frac{e^z + 1}{e^z - 1} \;=\; \frac{z}{2} \;+\; \frac{z}{e^z - 1}\,. \tag{6}$$

Pertanto tutto ciò che ci serve è uno sviluppo in serie di potenze della funzione $\frac{z}{e^z-1}$; si noti che tale funzione è definita e continua su tutto $\mathbb{R}$ (per $z=0$ si usi la serie di potenze della funzione esponenziale, o alternativamente la regola di de l'Hospital, che dà come valore 1). Scriviamo

$$\frac{z}{e^z-1} \;=:\; \sum_{n\geq 0} B_n \frac{z^n}{n!}\,. \tag{7}$$

I coefficienti B_n sono noti come i *numeri di Bernoulli*. Il termine sinistro di (6) è una funzione *pari* (ovvero, $f(z)=f(-z)$) e pertanto vediamo che $B_n=0$ per $n\geq 3$ dispari, mentre $B_1=-\frac{1}{2}$ corrisponde al termine di $\frac{z}{2}$ in (6).

Da

$$\Big(\sum_{n\geq 0} B_n \frac{z^n}{n!}\Big)(e^z-1) \;=\; \Big(\sum_{n\geq 0} B_n \frac{z^n}{n!}\Big)\Big(\sum_{n\geq 1} \frac{z^n}{n!}\Big) \;=\; z$$

otteniamo, confrontando i coefficienti di z^n:

$$\sum_{k=0}^{n-1} \frac{B_k}{k!(n-k)!} \;=\; \begin{cases} 1 & \text{per } n=1,\\ 0 & \text{per } n\neq 1. \end{cases} \tag{8}$$

Possiamo calcolare i numeri di Bernoulli ricorsivamente dalla (8). Il valore $n=1$ dà $B_0=1$, $n=2$ dà $\frac{B_0}{2}+B_1=0$, ovvero $B_1=-\frac{1}{2}$, e così via.

n	0	1	2	3	4	5	6	7	8
B_n	1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$

I primi numeri di Bernoulli

Ora abbiamo quasi finito: la combinazione di (6) e (7) dà

$$y\cot y \;=\; \sum_{k=0}^{\infty} B_{2k}\frac{(2iy)^{2k}}{(2k)!} \;=\; \sum_{k=0}^{\infty} \frac{(-1)^k 2^{2k} B_{2k}}{(2k)!} y^{2k},$$

ed ecco uscire, con (5), la formula di Eulero per $\zeta(2k)$:

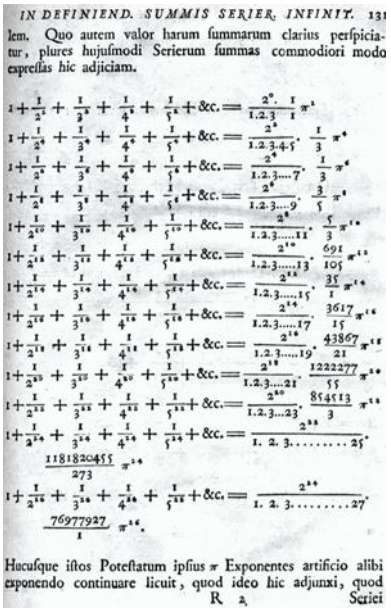
$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} \;=\; \frac{(-1)^{k-1}2^{2k-1}B_{2k}}{(2k)!}\,\pi^{2k} \qquad (k\in\mathbb{N}). \tag{9}$$

Esaminando la nostra tabella dei numeri di Bernoulli, otteniamo ancora una volta la somma $\sum \frac{1}{n^2}=\frac{\pi^2}{6}$ dal Capitolo 6, ed inoltre

$$\sum_{n=1}^{\infty}\frac{1}{n^4}=\frac{\pi^4}{90},\qquad \sum_{n=1}^{\infty}\frac{1}{n^6}=\frac{\pi^6}{945},\qquad \sum_{n=1}^{\infty}\frac{1}{n^8}=\frac{\pi^8}{9450},$$

$$\sum_{n=1}^{\infty}\frac{1}{n^{10}}=\frac{\pi^{10}}{93555},\qquad \sum_{n=1}^{\infty}\frac{1}{n^{12}}=\frac{691\,\pi^{12}}{638512875},\quad \cdots$$

Il numero di Bernoulli $B_{10}=\frac{5}{66}$ che ci porta a $\zeta(10)$ sembra innocuo, ma il valore successivo $B_{12}=-\frac{691}{2730}$, necessario per $\zeta(12)$, contiene il grande fattore primo 691 nel numeratore. Eulero calcolò per primo alcuni valori $\zeta(2k)$ senza notare la connessione con i numeri di Bernoulli. Solo l'apparizione dello strano numero primo 691 lo mise sulla buona traccia.



IN DEFINIEND. SUMMIS SERIER. INFINIT. 131

lem. Quo autem valor harum ſummarum clarius perſpicia-tur, plures hujuſmodi Serierum ſummas commodiori modo expreſſas hic adjiciam.

$$1+\tfrac{1}{2^2}+\tfrac{1}{3^2}+\tfrac{1}{4^2}+\tfrac{1}{5^2}+\&c. = \frac{2^0}{1.2.3}\,\pi^2$$
$$1+\tfrac{1}{2^4}+\tfrac{1}{3^4}+\tfrac{1}{4^4}+\tfrac{1}{5^4}+\&c. = \frac{2^3}{1.2.3.4.5}\cdot\frac{1}{3}\,\pi^4$$
$$1+\tfrac{1}{2^6}+\tfrac{1}{3^6}+\tfrac{1}{4^6}+\tfrac{1}{5^6}+\&c. = \frac{2^4}{1.2.3....7}\cdot\frac{1}{3}\,\pi^6$$
$$1+\tfrac{1}{2^8}+\tfrac{1}{3^8}+\tfrac{1}{4^8}+\tfrac{1}{5^8}+\&c. = \frac{2^7}{1.2.3....9}\cdot\frac{3}{5}\,\pi^8$$
$$1+\tfrac{1}{2^{10}}+\tfrac{1}{3^{10}}+\tfrac{1}{4^{10}}+\tfrac{1}{5^{10}}+\&c. = \frac{2^8}{1.2.3....11}\cdot\frac{5}{3}\,\pi^{10}$$
$$1+\tfrac{1}{2^{12}}+\tfrac{1}{3^{12}}+\tfrac{1}{4^{12}}+\tfrac{1}{5^{12}}+\&c. = \frac{2^{10}}{1.2.3....13}\cdot\frac{691}{105}\,\pi^{12}$$
$$1+\tfrac{1}{2^{14}}+\tfrac{1}{3^{14}}+\tfrac{1}{4^{14}}+\tfrac{1}{5^{14}}+\&c. = \frac{2^{12}}{1.2.3....15}\cdot\frac{35}{1}\,\pi^{14}$$
$$1+\tfrac{1}{2^{16}}+\tfrac{1}{3^{16}}+\tfrac{1}{4^{16}}+\tfrac{1}{5^{16}}+\&c. = \frac{2^{14}}{1.2.3....17}\cdot\frac{3617}{15}\,\pi^{16}$$
$$1+\tfrac{1}{2^{18}}+\tfrac{1}{3^{18}}+\tfrac{1}{4^{18}}+\tfrac{1}{5^{18}}+\&c. = \frac{2^{16}}{1.2.3....19}\cdot\frac{43867}{21}\,\pi^{18}$$
$$1+\tfrac{1}{2^{20}}+\tfrac{1}{3^{20}}+\tfrac{1}{4^{20}}+\tfrac{1}{5^{20}}+\&c. = \frac{2^{18}}{1.2.3....21}\cdot\frac{1222277}{55}\,\pi^{20}$$
$$1+\tfrac{1}{2^{22}}+\tfrac{1}{3^{22}}+\tfrac{1}{4^{22}}+\tfrac{1}{5^{22}}+\&c. = \frac{2^{20}}{1.2.3...23}\cdot\frac{854513}{3}\,\pi^{22}$$
$$1+\tfrac{1}{2^{24}}+\tfrac{1}{3^{24}}+\tfrac{1}{4^{24}}+\tfrac{1}{5^{24}}+\&c. = \frac{2^{22}}{1.2.3.\ldots\ldots25}\cdot\frac{1181820455}{273}\,\pi^{24}$$
$$1+\tfrac{1}{2^{26}}+\tfrac{1}{3^{26}}+\tfrac{1}{4^{26}}+\tfrac{1}{5^{26}}+\&c. = \frac{2^{24}}{1.2.3.\ldots\ldots27}\cdot\frac{76977927}{1}\,\pi^{26}.$$

Hucuſque iſtos Poteſtatum ipſius π Exponentes artificio alibi exponendo continuare licuit, quod ideo hic adjunxi, quod Seriei

R 2

La pagina 131 della "Introductio in Analysin Infinitorum" di Eulero nel 1748

Incidentalmente, poiché $\zeta(2k)$ converge a 1 per $k \longrightarrow \infty$, l'equazione (9) ci dice che i numeri $|B_{2k}|$ crescono molto rapidamente — qualcosa di non molto chiaro dai primi valori.

In contrasto con tutto ciò, si sa molto poco sui valori della funzione zeta di Riemann per gli interi dispari $k \geq 3$; si veda a pagina 48.

Bibliografia

[1] S. BOCHNER: *Book review of "Gesammelte Schriften" by Gustav Herglotz,* Bulletin Amer. Math. Soc. **1** (1979), 1020-1022.

[2] J. ELSTRODT: *Partialbruchzerlegung des Kotangens, Herglotz-Trick und die Weierstraßsche stetige, nirgends differenzierbare Funktion,* Math. Semesterberichte **45** (1998), 207-220.

[3] L. EULER: *Introductio in Analysin Infinitorum,* Tomus Primus, Lausanne 1748; Opera Omnia, Ser. 1, Vol. 8. In English: *Introduction to Analysis of the Infinite,* Book I (translated by J. D. Blanton), Springer-Verlag, New York 1988.

[4] L. EULER: *Institutiones calculi differentialis cum ejus usu in analysi finitorum ac doctrina serierum,* Petersburg 1755; Opera Omnia, Ser. 1, Vol. 10.

Il problema dell'ago di Buffon

Capitolo 21

Un nobile francese, Georges Louis Leclerc, Conte di Buffon, pose il seguente problema nel 1777:

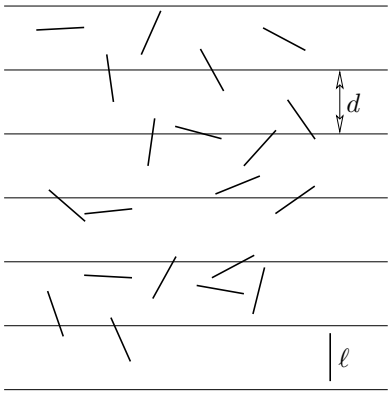
> *Si supponga di lasciar cadere un ago corto su di un foglio a righe. Qual è la probabilità che l'ago venga trovarsi in una posizione tale da incrociare una delle righe?*

Il Conte di Buffon

La probabilità dipende dalla distanza d fra le righe sul foglio e dipende dalla lunghezza ℓ dell'ago che facciamo cadere — o meglio, dipende soltanto dal rapporto $\frac{\ell}{d}$. Un ago *corto* nel nostro caso è un ago di lunghezza $\ell \leq d$. In altre parole, un ago corto è un ago che non può intersecare due righe allo stesso tempo (e verrà a toccare due righe soltanto con probabilità zero). La risposta al problema di Buffon può sorprendere: essa ha a che fare col numero π.

Teorema ("Il problema dell'ago di Buffon")
Se un ago corto, di lunghezza ℓ, è lasciato cadere su un foglio su cui sono tracciate delle righe equispaziate e la distanza fra due righe vicine è $d \geq \ell$, allora la probabilità che l'ago venga a trovarsi in una posizione in cui incroci una delle righe è esattamente

$$p \;=\; \frac{2}{\pi}\frac{\ell}{d}.$$



Tale risultato significa che da un esperimento si possono ottenere dei valori approssimativi per π: se lasciate cadere un ago N volte ed ottenete una risposta positiva (un'intersezione) in P casi, allora $\frac{P}{N}$ dovrebbe valere approssimativamente $\frac{2}{\pi}\frac{\ell}{d}$, ovvero, π dovrebbe essere approssimato da $\frac{2\ell N}{dP}$. Il test più esteso (ed esaustivo) fu forse quello effettuato da Lazzarini nel 1901, il quale pare avesse costruito addirittura una macchina per lasciar cadere un bastoncino 3408 volte (con $\frac{\ell}{d} = \frac{5}{6}$). Egli trovò che il bastoncino incrociò una delle righe 1808 volte, il che fornisce l'approssimazione $\pi \approx 2 \cdot \frac{5}{6}\,\frac{3408}{1808} = 3.1415929....$, la quale è corretta fino a sei cifre rispetto a π, e ben troppo precisa per essere vera! (I valori scelti da Lazzarini portano direttamente ad una ben nota approssimazione $\pi \approx \frac{355}{113}$; si veda a pagina 37. Questo spiega le scelte più che ambigue di 3408 e $\frac{5}{6}$, dove $\frac{5}{6}\,3408$ è un multiplo di 355. Si veda [5] per una discussione sull'inganno di Lazzarini.)

Il problema dell'ago può essere risolto calcolando un integrale. Lo faremo più avanti, e con questo metodo risolveremo anche il problema per un ago lungo. Tuttavia la *Book Proof*, presentata da E. Barbier nel 1860, non fa uso di integrali. Essa si limita a lasciar cadere un ago diverso...

Se si lascia cadere un ago *qualsiasi*, corto o lungo, allora il numero atteso di incroci sarà

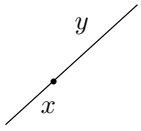
$$E \;=\; p_1 + 2p_2 + 3p_3 + \ldots,$$

dove p_1 è la probabilità che l'ago venga a cadere esattamente su un incrocio, p_2 la probabilità di ottenere esattamente due incroci, p_3 la probabilità di ottenere tre incroci, etc. La probabilità di avere almeno un incrocio, richiesta da Buffon, è pertanto

$$p \;=\; p_1 + p_2 + p_3 + \ldots$$

(Gli eventi in cui l'ago venga a cadere esattamente su una riga, o con un'estremità su una delle linee, hanno probabilità zero — pertanto possono essere ignorati nella nostra discussione.)

D'altro canto, se l'ago è *corto* allora la probabilità che vi sia più di un incrocio è zero, $p_2 = p_3 = \ldots = 0$, e pertanto abbiamo $E = p$: la probabilità che stiamo cercando è semplicemente il numero atteso di incroci. Questa riformulazione è estremamente utile, poiché ora possiamo usare la linearità del valore atteso (si veda a pagina 94). In effetti, scriviamo $E(\ell)$ per il valore atteso del numero di incroci che saranno prodotti lasciando cadere un ago dritto di lunghezza ℓ. Se tale lunghezza è $\ell = x + y$, e consideriamo la "parte anteriore" dell'ago di lunghezza x e la parte posteriore di lunghezza y separatamente, allora otteniamo



$$E(x+y) \;=\; E(x) + E(y),$$

poiché gli incroci prodotti sono sempre quelli prodotti dalla parte anteriore più quelli della parte posteriore.

Per induzione su n questa equazione funzionale implica che $E(nx) = nE(x)$ per ogni $n \in \mathbb{N}$, e quindi che $mE(\frac{n}{m}x) = E(m\frac{n}{m}x) = E(nx) = nE(x)$, pertanto $E(rx) = rE(x)$ vale per ogni $r \in \mathbb{Q}$ *razionale*. Inoltre, $E(x)$ è chiaramente monotona per $x \geq 0$, dal che otteniamo che $E(x) = cx$ per ogni $x \geq 0$, dove $c = E(1)$ è una costante.

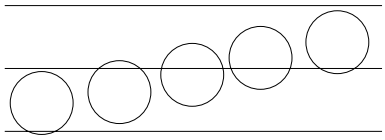
Ma quanto vale la costante?

Per scoprirlo useremo aghi di forma diversa. Facciamo dunque cadere un ago "poligonale" di lunghezza totale ℓ, che sia composto da diversi segmenti. Allora il numero di incroci che esso produce è (con probabilità 1) la somma del numero di incroci prodotto dai suoi segmenti. Pertanto, il numero atteso di incroci è di nuovo

$$E \;=\; c\,\ell,$$

per la linearità del valore atteso. (In effetti non è nemmeno importante che i suoi segmenti siano uniti in modo rigido o flessibile!)

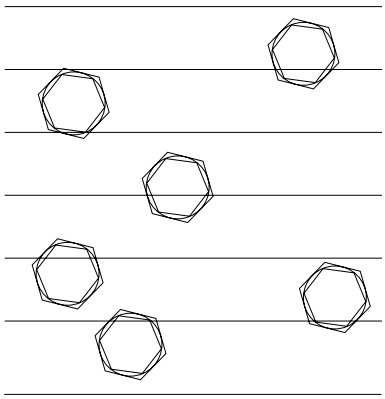
La chiave alla soluzione di Barbier del problema dell'ago di Buffon consiste nel considerare un ago che sia un cerchio perfetto C di diametro d, che ha lunghezza $x = d\pi$. Tale ago, se lasciato cadere su un foglio a righe, produce sempre esattamente due intersezioni.



Il cerchio può essere approssimato da poligoni. Immaginiamo soltanto di far cadere con l'ago circolare C un poligono inscritto P_n, e così anche un poligono circoscritto P^n. Ogni riga che intersechi P_n intersecherà anche C e se una riga interseca C allora essa toccherà anche P^n. Pertanto il valore atteso del numero di intersezioni soddisfa

$$E(P_n) \;\leq\; E(C) \;\leq\; E(P^n).$$

Ora sia P_n sia P^n sono poligoni, pertanto il numero di incroci che possiamo aspettarci è "c volte la lunghezza" per entrambi, mentre per C è 2, da cui



$$c\,\ell(P_n) \;\leq\; 2 \;\leq\; c\,\ell(P^n). \tag{1}$$

Sia P_n sia P^n approssimano C per $n \longrightarrow \infty$. In particolare,

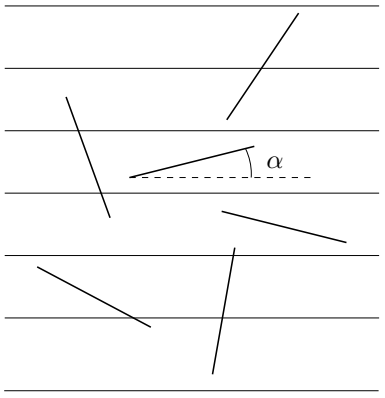
$$\lim_{n\to\infty} \ell(P_n) \;=\; d\pi \;=\; \lim_{n\to\infty} \ell(P^n),$$

e pertanto per $n \longrightarrow \infty$ deduciamo da (1) che

$$c\,d\pi \;\leq\; 2 \;\leq c\,d\pi,$$

il che dà $c = \frac{2}{\pi}\frac{1}{d}$. □

Ma *avremmo potuto* dimostrarlo anche con l'analisi! Il trucco per ottenere un integrale "facile" è di considerare in primo luogo la pendenza dell'ago; diciamo che sia lasciato cadere con un angolo α dalla retta orizzontale, dove α sia nell'intervallo $0 \leq \alpha \leq \frac{\pi}{2}$. (Ignoreremo il caso in cui l'ago venga a posizionarsi con una pendenza negativa, poiché tale caso è simmetrico al caso della pendenza positiva e produce la stessa probabilità.) Un ago con un angolo α ha lunghezza $\ell \sin\alpha$ e la probabilità che tale ago incroci una delle righe orizzontali di distanza d è $\frac{\ell \sin\alpha}{d}$. Pertanto otteniamo la probabilità facendo la media su tutti gli angoli α possibili,



$$p \;=\; \frac{2}{\pi}\int_0^{\pi/2} \frac{\ell\sin\alpha}{d}\,d\alpha \;=\; \frac{2}{\pi}\frac{\ell}{d}\big[-\cos\alpha\big]_0^{\pi/2} \;=\; \frac{2}{\pi}\frac{\ell}{d}.$$

Per un ago lungo, otteniamo la stessa probabilità $\frac{\ell\sin\alpha}{d}$ fintanto che $\ell\sin\alpha \leq d$, ovvero, nell'intervallo $0 \leq \alpha \leq \arcsin\frac{d}{\ell}$. Tuttavia, per angoli α più grandi l'ago *deve* incrociare una riga, dunque la probabilità è 1. Pertanto calcoliamo

$$\begin{aligned} p \;&=\; \frac{2}{\pi}\Big(\int_0^{\arcsin(d/\ell)} \frac{\ell\sin\alpha}{d}\,d\alpha \;+\; \int_{\arcsin(d/\ell)}^{\pi/2} 1\,d\alpha\Big) \\ &=\; \frac{2}{\pi}\Big(\frac{\ell}{d}\big[-\cos\alpha\big]_0^{\arcsin(d/\ell)} + \Big(\frac{\pi}{2} - \arcsin\frac{d}{\ell}\Big)\Big) \\ &=\; 1 + \frac{2}{\pi}\Big(\frac{\ell}{d}\Big(1 - \sqrt{1-\frac{d^2}{\ell^2}}\Big) - \arcsin\frac{d}{l}\Big) \end{aligned}$$

per $\ell \geq d$.

La risposta quindi non è ugualmente bella per un ago più lungo, ma ci fornisce un buon esercizio: si dimostri ("giusto per sicurezza") che la formula dà $\frac{2}{\pi}$ per $\ell = d$, che è strettamente crescente in ℓ, e che essa tende a 1 per $\ell \longrightarrow \infty$.

Bibliografia

[1] E. BARBIER: *Note sur le problème de l'aiguille et le jeu du joint couvert,* J. Mathématiques Pures et Appliquées (2) **5** (1860), 273-286.

[2] L. BERGGREN, J. BORWEIN & P. BORWEIN, EDS.: *Pi: A Source Book,* Springer-Verlag, New York 1997.

[3] G. L. LECLERC, COMTE DE BUFFON: *Essai d'arithmétique morale,* Appendix to "Histoire naturelle générale et particulière," Vol. 4, 1777.

[4] D. A. KLAIN & G.-C. ROTA: *Introduction to Geometric Probability,* "Lezioni Lincee," Cambridge University Press 1997.

[5] T. H. O'BEIRNE: *Puzzles and Paradoxes,* Oxford University Press, London 1965.

"Hai un problema?"

Calcolo Combinatorio

"Un malinconico quadro latino"

Il principio del casellario e la conta doppia

Capitolo 22

Alcuni principi matematici, come i due che compaiono nel titolo di questo capitolo, sono così evidenti che potreste pensare che producano soltanto risultati ovvi. Per convincervi che non è necessariamente così, li illustriamo con esempi che Paul Erdős suggerì di inserire nel *Libro*. Incontreremo esempi di questi principi anche nei capitoli successivi.

Il principio del casellario (NdT: Pigeon-hole principle nell'edizione inglese).
Se n oggetti sono disposti in r contenitori, con $r < n$, allora almeno uno dei contenitori contiene più di un oggetto.

"Il principio dal punto di vista dei piccioni"

Questo risultato è davvero ovvio, non vi è nulla da dimostrare. Nel linguaggio delle trasformazioni il nostro principio si traduce come segue: siano N e R due insiemi finiti tali che

$$|N| = n > r = |R|,$$

e sia $f : N \longrightarrow R$ una trasformazione. Allora esiste un $a \in R$ tale che $|f^{-1}(a)| \geq 2$. Possiamo persino affermare una disuguaglianza più forte: esiste un $a \in R$ tale che

$$|f^{-1}(a)| \ \geq \ \left\lceil \frac{n}{r} \right\rceil. \tag{1}$$

In effetti, in caso contrario avremmo $|f^{-1}(a)| < \frac{n}{r}$ per ogni a, e pertanto $n = \sum_{a \in R} |f^{-1}(a)| < r\,\frac{n}{r} = n$, il che non può essere.

1. Numeri

Proposizione. *Si considerino i numeri $1, 2, 3, \ldots, 2n$ e se ne scelgano $n + 1$ arbitrariamente. Allora due fra questi $n + 1$ numeri sono primi fra loro.*

Questo è di nuovo ovvio: devono esistere due numeri che distino fra loro solo di una unità e pertanto siano primi fra loro.

Ma proviamo ora a girare la condizione.

Proposizione. *Si supponga ancora $A \subseteq \{1, 2, \ldots, 2n\}$ con $|A| = n + 1$. Allora ci sono sempre due numeri in A tali che uno divida l'altro.*

Entrambi i risultati non sono più validi se si sostituisce $n+1$ con n: per questo si considerino gli insiemi $\{2, 4, 6, \ldots, 2n\}$, rispettivamente $\{n+1, n+2, \ldots, 2n\}$

Questo non è così chiaro. Come ci disse Erdős, egli pose questa domanda al giovane Lajos Pośa durante una cena e alla fine del pasto Lajos aveva la risposta. Il quesito è rimasto una delle domande "di iniziazione" alla matematica preferite da Erdős. La soluzione è fornita dal principio del casellario. Si scriva ogni numero $a \in A$ nella forma $a = 2^k m$, dove m è un numero dispari compreso fra 1 e $2n-1$. Poiché vi sono $n+1$ numeri in A, ma solo n componenti dispari diverse, devono esserci due numeri in A con la *stessa* componente dispari. Pertanto uno è multiplo dell'altro. □

2. Successioni

Questo è un altro fra i quesiti preferiti da Erdős, contenuto in una pubblicazione di Erdős e Szekeres sui problemi di Ramsey.

Proposizione. *In ogni successione $a_1, a_2, \ldots, a_{mn+1}$ di $mn+1$ numeri reali distinti, esiste una sottosuccessione crescente*

$$a_{i_1} < a_{i_2} < \ldots < a_{i_{m+1}} \qquad (i_1 < i_2 < \ldots < i_{m+1})$$

di lunghezza $m+1$, o una sottosuccessione decrescente

$$a_{j_1} > a_{j_2} > \ldots > a_{j_{n+1}} \qquad (j_1 < j_2 < \ldots < j_{n+1})$$

di lunghezza $n+1$, oppure entrambe.

Questa volta l'applicazione del principio del casellario non è immediata. Si associ ad ogni a_i il numero t_i che è la lunghezza della *più lunga successione crescente* che inizia con a_i. Se $t_i \geq m+1$ per un i dato, allora abbiamo una sottosuccessione crescente di lunghezza $m+1$. Si supponga allora che $t_i \leq m$ per ogni i. La funzione $f : a_i \longmapsto t_i$ che trasforma $\{a_1, \ldots, a_{mn+1}\}$ in $\{1, \ldots, m\}$ ci dice per (1) che vi è un $s \in \{1, \ldots, m\}$ tale che $f(a_i) = s$ per $\frac{mn}{m}+1 = n+1$ numeri a_i. Siano $a_{j_1}, a_{j_2}, \ldots, a_{j_{n+1}}$ $(j_1 < \ldots < j_{n+1})$ tali numeri. Ora guardiamo due numeri consecutivi a_{j_i}, $a_{j_{i+1}}$. Se $a_{j_i} < a_{j_{i+1}}$, allora otterremmo una sottosuccessione crescente di lunghezza s che inizia con $a_{j_{i+1}}$, e quindi una sottosuccessione crescente di lunghezza $s+1$ che inizia con a_{j_i}, il che non può essere dal momento che $f(a_{j_i}) = s$. Otteniamo pertanto una sottosuccessione decrescente $a_{j_1} > a_{j_2} > \ldots > a_{j_{n+1}}$ di lunghezza $n+1$. □

Il lettore può divertirsi a dimostrare che per mn numeri l'affermazione non è più vera in generale.

Questo risultato dall'aria semplice sulle sottosuccessioni monotone ha una conseguenza assolutamente non ovvia sul concetto di *dimensione di un grafo*. Non abbiamo bisogno qui della nozione di dimensione per i grafi in generale, ma solo per grafi completi K_n. Essa si può formulare nel seguente modo. Sia $N = \{1, \ldots, n\}$, $n \geq 3$ e si considerino m permutazioni $\pi_1, \ldots, \pi_m$ di N. Diciamo che le permutazioni π_i *rappresentano* K_n se per ogni tripletta di numeri distinti i, j, k esiste una permutazione π in cui k appare *dopo* sia i che j. La dimensione di K_n è allora il più piccolo m per cui esista una rappresentazione $\pi_1, \ldots, \pi_m$.

Ad esempio abbiamo $\dim(K_3) = 3$ dal momento che ognuno degli ultimi tre numeri deve apparire per ultimo, come in $\pi_1 = (1,2,3)$, $\pi_2 = (2,3,1)$, $\pi_3 = (3,1,2)$. Ma cosa succede con K_4? Si noti dapprima che $\dim(K_n) \le \dim(K_{n+1})$: basta eliminare $n+1$ in una rappresentazione di K_{n+1}. Quindi, $\dim(K_4) \ge 3$ e, in effetti, $\dim(K_4) = 3$, prendendo

$$\pi_1 = (1,2,3,4), \quad \pi_2 = (2,4,3,1), \quad \pi_3 = (1,4,3,2).$$

Non è così semplice dimostrare che $\dim(K_5) = 4$, ma a questo punto, sorprendentemente, la dimensione resta ferma a 4 fino a $n = 12$, mentre $\dim(K_{13}) = 5$. Allora $\dim(K_n)$ sembra una funzione piuttosto "selvaggia". Ma, non lo è! In effetti, quando n tende all'infinito, $\dim(K_n)$ è una funzione molto tranquilla — e la chiave per trovarne una limitazione inferiore è il principio del casellario. Affermiamo che

$$\dim(K_n) \;\ge\; \log_2 \log_2 n. \tag{2}$$

Poiché, come abbiamo visto, $\dim(K_n)$ è una funzione monotona in n, basta verificare (2) per $n = 2^{2^p} + 1$, ovvero, dobbiamo dimostrare che

$$\dim(K_n) \;\ge\; p+1 \qquad \text{per} \quad n = 2^{2^p} + 1.$$

Supponiamo, al contrario, che $\dim(K_n) \le p$, e siano $\pi_1, \ldots, \pi_p$ permutazioni che rappresentano $N = \{1, 2, \ldots, 2^{2^p} + 1\}$. Ora usiamo il nostro risultato sulle sottosuccessioni monotone p volte. In π_1 esiste una sottosuccessione monotona A_1 di lunghezza $2^{2^{p-1}} + 1$ (non importa se crescente o decrescente). Consideriamo questo insieme A_1 in π_2. Usando nuovamente il nostro risultato, troviamo una sottosuccessione monotona A_2 di A_1 in π_2 di lunghezza $2^{2^{p-2}} + 1$ e A_2 è, ovviamente, anch'essa monotona in π_1. Continuando, troviamo infine una sottosuccessione A_p di dimensione $2^{2^0} + 1 = 3$ monotona in *tutte* le permutazioni π_i. Sia $A_p = (a,b,c)$; allora o $a < b < c$ oppure $a > b > c$ *per ogni* π_i. Ma questo non può essere, poiché deve esserci una permutazione in cui b appaia dopo a e c. □

La corretta crescita asintotica fu fornita da Joel Spencer (la maggiorazione) e da Füredi, Hajnal, Rödl e Trotter (la minorazione):

$$\dim(K_n) \;=\; \log_2 \log_2 n + (\frac{1}{2} + o(1)) \log_2 \log_2 \log_2 n.$$

Ma la storia non finisce qui: molto recentemente, Morris e Hoşten hanno trovato un metodo che, in linea di principio, stabilisce il valore *preciso* di $\dim(K_n)$. Usando il loro risultato ed il calcolatore si possono ottenere i valori riportati a margine. Questo è veramente sbalorditivo! Si consideri soltanto quante permutazioni di dimensione 1422564 esistano. Come si fa a decidere se ne servano 7 o 8 per rappresentare $K_{1422564}$?

π_1:	1	2	3	5	6	7	8	9	10	11	12	4
π_2:	2	3	4	8	7	6	5	12	11	10	9	1
π_3:	3	4	1	11	12	9	10	6	5	8	7	2
π_4:	4	1	2	10	9	12	11	7	8	5	6	3

Queste quattro permutazioni rappresentano K_{12}

$$\begin{aligned}
\dim(K_n) \le 4 &\iff n \le 12\\
\dim(K_n) \le 5 &\iff n \le 81\\
\dim(K_n) \le 6 &\iff n \le 2646\\
\dim(K_n) \le 7 &\iff n \le 1422564
\end{aligned}$$

3. Somme

Paul Erdős attribuisce a Andrew Vázsonyi e Marta Sved la bella applicazione seguente del principio del casellario:

Proposizione. *Si considerino n interi arbitrari $a_1, \ldots, a_n$, non necessariamente distinti. Allora vi è sempre un insieme di numeri consecutivi $a_{k+1}, a_{k+2}, \ldots, a_\ell$ la cui somma $\sum_{i=k+1}^{\ell} a_i$ sia un multiplo di n.*

Per la dimostrazione, poniamo $N = \{0, 1, \ldots, n\}$ e $R = \{0, 1, \ldots, n-1\}$. Si consideri la trasformazione $f : N \to R$, dove $f(m)$ è il resto di $a_1 + \ldots + a_m$ dopo la divisione per n. Poiché $|N| = n+1 > n = |R|$, ne segue che vi sono due somme $a_1 + \ldots + a_k$, $a_1 + \ldots + a_\ell$ $(k < \ell)$ con lo *stesso* resto, dove la prima somma può essere la somma vuota indicata con 0. Ne segue che

$$\sum_{i=k+1}^{\ell} a_i \;=\; \sum_{i=1}^{\ell} a_i \;-\; \sum_{i=1}^{k} a_i$$

ha resto 0, il che conclude la dimostrazione. □

Affrontiamo adesso il secondo principio: la conta doppia.

Conta doppia.
Si considerino due insiemi finiti R e C ed un sottoinsieme $S \subseteq R \times C$. Ogniqualvolta $(p, q) \in S$, diciamo che p e q sono incidenti. Se r_p indica il numero di elementi incidenti con $p \in R$, e c_q denota il numero di elementi incidenti con $q \in C$, allora

$$\sum_{p \in R} r_p \;=\; |S| \;=\; \sum_{q \in C} c_q. \tag{3}$$

Ancora, non vi è nulla da dimostrare. La prima somma classifica la coppie in S in base al primo elemento, mentre la seconda somma classifica le stessse coppie in base al secondo elemento.

C'è un modo utile di rappresentare l'insieme S. Si consideri la *matrice d'incidenza* $A = (a_{pq})$ di S, le cui righe e colonne sono ordinate in base agli elementi di R e C, rispettivamente, con

$$a_{pq} = \begin{cases} 1 & \text{se } (p, q) \in S \\ 0 & \text{se } (p, q) \notin S. \end{cases}$$

Con questa configurazione, r_p è la somma della p-esima riga di A e c_q è la somma della q-esima colonna. Pertanto la prima somma in (3) addiziona gli elementi di A (ovvero, conta gli elementi in S) riga per riga, e la seconda somma colonna per colonna.

L'esempio che segue dovrebbe chiarire questa corrispondenza. Sia $R = C = \{1, 2, \ldots, 8\}$ e si ponga $S = \{(i, j) : i \text{ divide } j\}$. Otteniamo allora la matrice riportata a margine, che mostra solo gli 1.

R \ C	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2		1		1		1		1
3			1			1		
4				1				1
5					1			
6						1		
7							1	
8								1

4. Ancora numeri

Si guardi la tabella a margine. Il numero di 1 nella colonna j è precisamente il numero di divisori di j; indichiamo questo numero con $t(j)$. Chiediamoci quanto è grande questo numero $t(j)$ *in media* quando j varia tra 1 e n. Pertanto, studiamo la quantità

$$\bar{t}(n) = \frac{1}{n}\sum_{j=1}^{n} t(j).$$

n	1	2	3	4	5	6	7	8
$\bar{t}(n)$	1	$\frac{3}{2}$	$\frac{5}{3}$	2	2	$\frac{7}{3}$	$\frac{16}{7}$	$\frac{5}{2}$

I primi valori di $\bar{t}(n)$

Quanto è grande $\bar{t}(n)$ per un n arbitrario? A prima vista, questo calcolo sembra senza speranza. Per i numeri primi p abbiamo $t(p) = 2$, mentre per 2^k otteniamo un grande numero $t(2^k) = k+1$. Quindi, $t(n)$ è una funzione "selvaggia" a salti e possiamo supporre che lo stesso succeda per $\bar{t}(n)$. In realtà è vero il contrario! La conta in due sensi dà una risposta semplice ed inaspettata.

Si consideri ancora la matrice A precedentemente introdotta per gli interi da 1 a n. Contando colonna per colonna otteniamo $\sum_{j=1}^{n} t(j)$. Quanti 1 ci sono nella riga i-esima? Facile, gli 1 corrispondono ai multipli di i: $1i, 2i, \ldots$ e l'ultimo multiplo non superiore a n è $\lfloor \frac{n}{i} \rfloor i$. Pertanto otteniamo

$$\bar{t}(n) \;=\; \frac{1}{n}\sum_{j=1}^{n} t(j) \;=\; \frac{1}{n}\sum_{i=1}^{n} \left\lfloor \frac{n}{i} \right\rfloor \;\leq\; \frac{1}{n}\sum_{i=1}^{n} \frac{n}{i} \;=\; \sum_{i=1}^{n} \frac{1}{i},$$

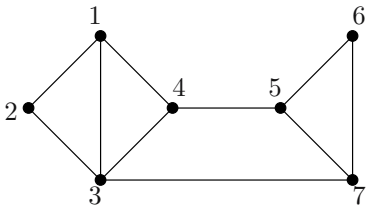
dove in ogni addendo l'errore nel passare da $\lfloor \frac{n}{i} \rfloor$ a $\frac{n}{i}$ è inferiore a 1. Ora l'ultima somma è l'ennesimo numero armonico H_n, così otteniamo $H_n - 1 < \bar{t}(n) < H_n$ che insieme alle stime di H_n a pagina 11 dà

$$\log n - 1 \;<\; H_n - 1 - \frac{1}{n} \;<\; \bar{t}(n) \;<\; H_n \;<\; \log n + 1.$$

Abbiamo pertanto dimostrato il notevole risultato che, mentre $t(n)$ è totalmente erratico, la media $\bar{t}(n)$ si comporta benissimo: essa differisce da $\log n$ per meno di 1.

5. Grafi

Sia G un grafo semplice finito con insieme di vertici V ed insieme di spigoli E. Abbiamo definito nel Capitolo 11 il *grado* $d(v)$ di un vertice v come il numero di spigoli aventi v come estremità. Nell'esempio della figura a fianco, i vertici $1, 2, \ldots, 7$ hanno grado $3, 2, 4, 3, 3, 2, 3$, rispettivamente.



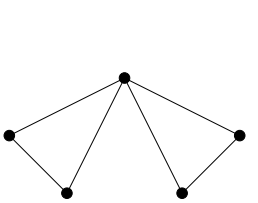
Quasi ogni libro sulla teoria dei grafi inizia con il seguente risultato (che abbiamo già incontrato nei Capitoli 11 e 17):

$$\sum_{v \in V} d(v) \;=\; 2|E|. \tag{4}$$

Per la dimostrazione si consideri $S \subseteq V \times E$, dove S è l'insieme delle coppie (v, e) tali che $v \in V$ sia un vertice all'estremità di $e \in E$. Contando

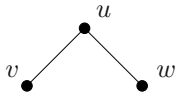
S in due sensi si ha da un lato $\sum_{v\in V} d(v)$, poiché ogni vertice contribuisce con $d(v)$ alla conta, e dall'altro $2|E|$, poiché ogni spigolo ha due estremità. □

Per quanto appaia semplice, il risultato (4) ha molte conseguenze importanti, alcune delle quali saranno discusse mentre procediamo. Vogliamo isolare in questa sezione la bellissima applicazione ad un *problema di massimo* sui grafi. Ecco il problema:

> *Si supponga che $G = (V, E)$ abbia n vertici e non contenga cicli di lunghezza 4 (indichiamo con C_4 questa condizione) ovvero nessun sottografo*  . *Quanti spigoli può avere al massimo G ?*

Ad esempio, il grafo a margine con 5 vertici non contiene 4-cicli ed ha 6 spigoli. Il lettore può facilmente dimostrare che su 5 vertici il numero massimo di spigoli è 6 e che questo grafo è in effetti l'unico grafo con 5 vertici e con 6 spigoli a non contenere 4-cicli.

Affrontiamo il problema generale. Sia G un grafo con n vertici senza 4-cicli. Come sopra indichiamo con $d(u)$ il grado di u. Ora contiamo in due modi l'insieme S delle coppie $(u, \{v, w\})$ dove u è adiacente a v ed a w, con $v \neq w$. In altre parole, contiamo tutte le presenze di sottografi del tipo



Sommando su u, troviamo $|S| = \sum_{u\in V} \binom{d(u)}{2}$. D'altro canto, ogni coppia $\{v, w\}$ ha al massimo un vicino in comune (secondo la condizione C_4). Pertanto $|S| \leq \binom{n}{2}$ e concludiamo

$$\sum_{u\in V} \binom{d(u)}{2} \;\leq\; \binom{n}{2},$$

ovvero

$$\sum_{u\in V} d(u)^2 \;\leq\; n(n-1) + \sum_{u\in V} d(u). \tag{5}$$

Inoltre (il che è alquanto tipico per questo genere di problemi di estremo) applichiamo la disuguaglianza di Cauchy-Schwarz ai vettori $(d(u_1), \ldots, d(u_n))$ e $(1, 1, \ldots, 1)$, ottenendo

$$\Big(\sum_{u\in V} d(u)\Big)^2 \;\leq\; n \sum_{u\in V} d(u)^2,$$

e pertanto in base a (5)

$$\Big(\sum_{u\in V} d(u)\Big)^2 \;\leq\; n^2(n-1) + n \sum_{u\in V} d(u).$$

Usando (4) troviamo

$$4\,|E|^2 \;\leq\; n^2(n-1) + 2n\,|E|,$$

ovvero

$$|E|^2 - \frac{n}{2}|E| - \frac{n^2(n-1)}{4} \;\leq\; 0.$$

Risolvendo l'equazione quadratica corrispondente otteniamo pertanto il seguente risultato di Istvan Reiman.

Teorema. *Se il grafo G a n vertici non contiene 4-cicli, allora*

$$|E| \;\leq\; \left\lfloor \frac{n}{4}\left(1+\sqrt{4n-3}\right)\right\rfloor. \tag{6}$$

Per $n = 5$ si ottiene $|E| \leq 6$ ed il grafo dell'esempio mostra che l'uguaglianza può aver luogo.

La conta doppia ha pertanto prodotto in un modo semplice una maggiorazione per il numero di spigoli. Ma quanto è fine la maggiorazione (6) in generale? Il seguente splendido esempio [2] [3] [6] mostra che essa è quasi ottimale. Come accade spesso in questi problemi, è la geometria finita ad indicare la strada.

Presentando l'esempio supponiamo che il lettore abbia familiarità con il concetto di campo finito $\mathbb{Z}_p$ di interi modulo un numero primo p (si veda a pagina 20). Si consideri lo spazio vettoriale tridimensionale X su $\mathbb{Z}_p$. Costruiamo da X il grafo seguente G_p. I vertici di G_p sono i sottospazi monodimensionali $[\boldsymbol{v}] := \mathrm{span}_{\mathbb{Z}_p}\{\boldsymbol{v}\}$, $\mathbf{0} \neq \boldsymbol{v} \in X$; connettiamo due di tali sottospazi $[\boldsymbol{v}]$, $[\boldsymbol{w}]$ con uno spigolo se

$$\langle \boldsymbol{v}, \boldsymbol{w}\rangle \;=\; v_1w_1 + v_2w_2 + v_3w_3 \;=\; 0.$$

Si noti che non importa quale vettore $\neq \mathbf{0}$ scegliamo nel sottospazio. Nel linguaggio della geometria, i vertici sono i *punti* del piano di proiezione su $\mathbb{Z}_p$, e $[\boldsymbol{w}]$ è adiacente a $[\boldsymbol{v}]$ se $\boldsymbol{w}$ giace sulla *retta polare* di $\boldsymbol{v}$.

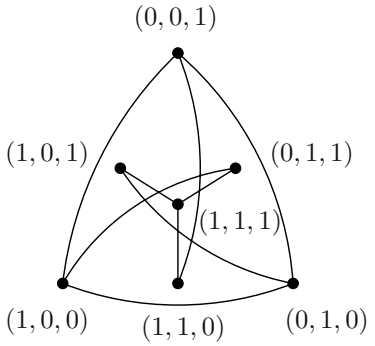
Ad esempio, il grafo G_2 non ha 4-cicli e contiene 9 spigoli, il che raggiunge quasi il limite superiore di 10 dato da (6). Vogliamo mostrare che questo è vero per ogni numero primo p.

Dimostriamo dapprima che G_p soddisfa la condizione C_4. Se $[\boldsymbol{u}]$ è un vicino comune di $[\boldsymbol{v}]$ e $[\boldsymbol{w}]$, allora $\boldsymbol{u}$ è una soluzione del sistema lineare

$$\begin{aligned} v_1x + \;v_2y + \;v_3z \;&= 0\\ w_1x + w_2y + w_3z \;&= 0. \end{aligned}$$

Poiché $\boldsymbol{v}$ e $\boldsymbol{w}$ sono linearmente indipendenti, deduciamo che lo spazio delle soluzioni ha dimensione 1 e pertanto che il vicino comune $[\boldsymbol{u}]$ è unico.

In seguito, ci chiediamo quanti vertici abbia G_p. È di nuovo la conta doppia che ci conduce al risultato. Lo spazio X contiene $p^3 - 1$ vettori $\neq \mathbf{0}$. Dal momento che ogni sottospazio monodimensionale contiene $p-1$ vettori $\neq \mathbf{0}$, deduciamo che X ha $\frac{p^3-1}{p-1} = p^2+p+1$ sottospazi monodimensionali, ovvero, G_p ha $n = p^2+p+1$ vertici. Analogamente, ogni sottospazio bidimensionale contiene $p^2 - 1$ vettori$\neq \mathbf{0}$, e pertanto $\frac{p^2-1}{p-1} = p+1$ sottospazi monodimensionali.



Il grafo G_2: i suoi vertici sono tutte le sette triplette non nulle (x, y, z)

Rimane da determinare il numero degli spigoli in G_p ovvero i gradi il che è lo stesso per via di (4). Per costruzione di G_p, i vertici adiacenti a $[\boldsymbol{u}]$ sono le soluzioni dell'equazione

$$u_1x \ + \ u_2y \ + \ u_3z \ = \ 0. \tag{7}$$

Lo spazio delle soluzioni di (7) è un sottospazio bidimensionale, pertanto vi sono $p+1$ vertici adiacenti a $[\boldsymbol{u}]$. Ma attenzione, può accadere che lo stesso $\boldsymbol{u}$ sia una soluzione di (7). In questo caso ci sono solo p vertici adiacenti a $[\boldsymbol{u}]$.

Riassumendo, otteniamo il risultato seguente: se $\boldsymbol{u}$ giace sulla *conica* data da $x^2+y^2+z^2=0$, allora $d([\boldsymbol{u}])=p$, altrimenti $d([\boldsymbol{u}])=p+1$. Rimane quindi da trovare il numero di sottospazi monodimensionali sulla conica

$$x^2+y^2+z^2=0.$$

Anticipiamo il risultato che dimostreremo fra un momento.

> **Proposizione.** *Vi sono esattamente p^2 soluzioni (x,y,z) dell'equazione $x^2+y^2+z^2=0$; di conseguenza (eliminando la soluzione nulla) G_p ha esattamente $\frac{p^2-1}{p-1}=p+1$ vertici di grado p.*

Con questo risultato, completiamo la nostra analisi di G_p. Vi sono $p+1$ vertici di grado p, dunque $(p^2+p+1)-(p+1)=p^2$ vertici di grado $p+1$. Usando (4), otteniamo

$$\begin{aligned} |E| &= \frac{(p+1)p}{2}+\frac{p^2(p+1)}{2} \ = \ \frac{(p+1)^2p}{2} \\ &= \frac{(p+1)p}{4}\,(1+(2p+1)) \ = \ \frac{p^2+p}{4}\,(1+\sqrt{4p^2+4p+1}). \end{aligned}$$

Ponendo $n=p^2+p+1$, la stessa equazione si legge

$$|E|=\frac{n-1}{4}\,(1+\sqrt{4n-3}),$$

e vediamo che questo è quasi in accordo con (6).

Dimostriamo ora la proposizione. L'argomentazione seguente è una splendida applicazione dell'algebra lineare che riguarda le matrici simmetriche ed i loro autovalori. Ritroveremo lo stesso metodo nel Capitolo 34, il che non è una coincidenza: entrambe le dimostrazioni provengono dalla stessa pubblicazione di Erdős, Rényi e Sós.

Rappresentiamo i sottospazi monodimensionali di X come prima mediante i vettori $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{p^2+p+1}$; essi sono linearmente indipendenti a due a due. Analogamente, possiamo rappresentare i sottospazi bidimensionali con lo *stesso* insieme di vettori, dove il sottospazio corrispondente a $\boldsymbol{u}=(u_1,u_2,u_3)$ è l'insieme delle soluzioni dell'equazione $u_1x+u_2y+u_3z=0$ come in (7). (Questo è semplicemente il principio della dualità dell'algebra

lineare.) Pertanto, secondo (7), un sottospazio monodimensionale rappresentato da $\boldsymbol{v}_i$ è contenuto nel sottospazio bidimensionale rappresentato da $\boldsymbol{v}_j$ se e solo se $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0$.

Si consideri ora la matrice $A = (a_{ij})$ di dimensioni $(p^2 + p + 1) \times (p^2 + p + 1)$, definita come segue: le righe e colonne di A corrispondono a $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p^2+p+1}$ (adottiamo la stessa numerazione per le righe e le colonne) tale che

$$a_{ij} := \begin{cases} 1 & \text{se } \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0, \\ 0 & \text{altrimenti.} \end{cases}$$

A è dunque una matrice reale simmetrica ed abbiamo $a_{ii} = 1$ se $\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle = 0$ ovvero precisamente quando $\boldsymbol{v}_i$ giace sulla conica $x^2 + y^2 + z^2 = 0$. Pertanto, tutto ciò che resta da dimostrare è che

$$traccia(A) \;=\; p + 1.$$

Dall'algebra lineare sappiamo che la traccia è uguale alla somma degli autovalori. E qui viene il trucco: A appare difficile, mentre la matrice A^2 è semplice da analizzare. Notiamo due proprietà:

- Ogni riga di A contiene 1 esattamente $p + 1$ volte. Ciò implica che $p + 1$ è un autovalore di A, poiché $A\mathbf{1} = (p + 1)\mathbf{1}$, dove $\mathbf{1}$ è il vettore degli 1.
- Per ogni coppia di righe distinte $\boldsymbol{v}_i, \boldsymbol{v}_j$ vi è esattamente una colonna con un 1 in entrambe le righe (la colonna corrispondente all'unico sottospazio generato da $\boldsymbol{v}_i, \boldsymbol{v}_j$).

Grazie a queste proprietà troviamo

$$A^2 = \begin{pmatrix} p+1 & 1 & \cdots & 1 \\ 1 & p+1 & & \vdots \\ \vdots & & \ddots & \\ 1 & \cdots & & p+1 \end{pmatrix} \;=\; pI \;+\; J,$$

dove I è la matrice identità e J è la matrice di tutti 1. Ora, J ha come autovalori $p^2 + p + 1$ (di molteplicità 1) e 0 (di molteplicità $p^2 + p$). Pertanto A^2 ha come autovalori $p^2 + 2p + 1 = (p + 1)^2$ di molteplicità 1 e p di molteplicità $p^2 + p$. Poiché A è reale e simmetrica, e dunque diagonalizzabile, troviamo che A ammette l'autovalore $p + 1$ o $-(p + 1)$ e $p^2 + p$ autovalori $\pm\sqrt{p}$. Dal primo fatto sopracitato, il primo autovalore deve essere $p + 1$. Supponiamo che $\sqrt{p}$ abbia molteplicità r, e $-\sqrt{p}$ molteplicità s, allora

$$traccia(A) \;=\; (p + 1) + r\sqrt{p} - s\sqrt{p}.$$

Ma ora siamo a casa: poiché la traccia è un intero, dobbiamo avere $r = s$, quindi $traccia(A) = p + 1$. □

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

La matrice per G_2

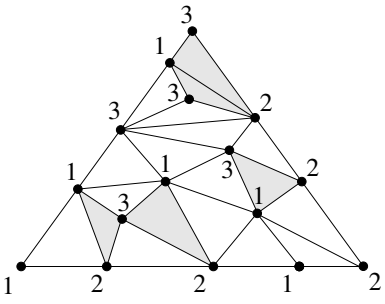
6. Il lemma di Sperner

Nel 1912, Luitzen Brouwer pubblicò il suo celebre teorema sui punti fissi:

> *Ogni funzione continua $f\colon B^n \longrightarrow B^n$ di una palla n-dimensionale in se stessa ammette un punto fisso (un punto $\boldsymbol{x} \in B^n$ tale che $f(\boldsymbol{x}) = \boldsymbol{x}$).*

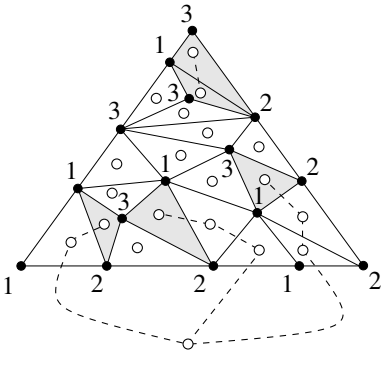
Per la dimensione 1, ovvero per un intervallo, questo segue facilmente dal teorema dei valori intermedi, ma per dimensioni più alte la dimostrazione di Brouwer necessitava di un macchinario sofisticato. Fu quindi una vera sorpresa quando nel 1928 il giovane Emanuel Sperner (aveva 23 anni all'epoca) produsse un semplice risultato combinatorio dal quale si potevano dedurre il teorema del punto fisso di Brouwer e l'invarianza della dimensione per trasformazioni continue biiettive. Inoltre l'ingegnoso lemma di Sperner è abbinato ad una altrettanto bella dimostrazione che è proprio la conta doppia.

Discutiamo il lemma di Sperner, e di conseguenza il teorema di Brouwer, per il primo caso interessante, quello della dimensione $n = 2$. Il lettore non dovrebbe avere difficoltà nell'estendere le dimostrazioni a dimensioni più alte (procedendo per induzione sulla dimensione).



I triangoli con tre diversi colori sono ombreggiati

Lemma di Sperner
Supponiamo che un "grande" triangolo con vertici V_1, V_2, V_3 sia triangolato (ovvero, scomposto in un numero finito di triangoli "piccoli" che condividano fra loro interi lati).
Supponiamo che ai vertici della triangolazione siano assegnati "colori" scelti nell'insieme $\{1, 2, 3\}$ in modo tale che V_i riceva il colore i (per ogni i) e che solo i colori i e j siano usati per i vertici che si trovano lungo lo spigolo da V_i a V_j (per $i \neq j$), mentre i vertici interni siano colorati arbitrariamente con 1, 2 o 3.
Allora nella triangolazione deve esistere un triangolo piccolo "tricolore", ovvero che abbia i tre vertici di tre diversi colori.



■ **Dimostrazione.** Dimostreremo un'affermazione più forte: il numero di triangoli tricolori non solo non è nullo, ma è sempre *dispari*.

Si consideri il grafo duale alla triangolazione, ma non si prendano tutti i suoi vertici — solo quelli che incrociano uno spigolo che abbia vertici con i colori (distinti) 1 e 2. In questo modo otteniamo un "grafo duale parziale" che ha grado 1 per tutti i vertici che corrispondono a triangoli tricolori, grado 2 per tutti i triangoli in cui appaiono i due colori 1 e 2 e grado 0 per i triangoli che non hanno nessuno dei colori 1 e 2. Pertanto solo i triangoli tricolori corrispondono a vertici di grado dispari (di grado 1).

Tuttavia, il vertice del grafo duale che corrisponde all'esterno della triangolazione ha grado dispari: in effetti, lungo lo spigolo grande da V_1 a V_2, vi è un numero dispari di cambi tra 1 e 2. Dunque un numero dispari di spigoli del grafo duale parziale incrocia questo spigolo grande, mentre gli altri spigoli grandi non possono avere sia 1 sia 2 come colori.

Ora poiché il numero di vertici dispari in ogni grafo finito è pari (secondo l'equazione (4)), troviamo che il numero di triangoli piccoli con tre diversi colori (corrispondenti a vertici interni dispari del nostro grafo duale) è dispari. □

Grazie a questo lemma è facile dimostrare il teorema di Brouwer.

■ **Dimostrazione del teorema del punto fisso di Brouwer (per $n = 2$).** Sia Δ il triangolo in $\mathbb{R}^3$ con vertici $\boldsymbol{e}_1 = (1,0,0)$, $\boldsymbol{e}_2 = (0,1,0)$, e $\boldsymbol{e}_3 = (0,0,1)$. Basta dimostrare che ogni trasformazione continua $f\colon \Delta \longrightarrow \Delta$ ha un punto fisso, poiché Δ è omeomorfa alla palla bidimensionale B_2.

Usiamo $\delta(\mathcal{T})$ per indicare la lunghezza massima di uno spigolo nella triangolazione $\mathcal{T}$. Si può facilmente costruire una successione infinita di triangolazioni $\mathcal{T}_1, \mathcal{T}_2, \ldots$ di Δ tale che la successione di diametri massimi $\delta(\mathcal{T}_k)$ converga a 0. Tale successione si può ottenere per costruzione esplicita, oppure induttivamente, ad esempio prendendo $\mathcal{T}_{k+1}$ come suddivisione baricentrica di $\mathcal{T}_k$.

Per ognuna di queste triangolazioni, definiamo una 3-colorazione dei suoi vertici $\boldsymbol{v}$ ponendo $\lambda(\boldsymbol{v}) := \min\{i : f(\boldsymbol{v})_i < v_i\}$, ovvero, $\lambda(\boldsymbol{v})$ è il più piccolo indice i tale che la coordinata i-esima di $f(\boldsymbol{v}) - \boldsymbol{v}$ sia negativa. Supponendo che f non abbia punti fissi, questo è ben definito. Per constatarlo, si noti che ogni $\boldsymbol{v} \in \Delta$ giace nel piano $x_1 + x_2 + x_3 = 1$ e dunque $\sum_i v_i = 1$. Pertanto se $f(\boldsymbol{v}) \neq \boldsymbol{v}$, allora almeno una delle coordinate di $f(\boldsymbol{v}) - \boldsymbol{v}$ deve essere negativa (ed almeno una deve essere positiva).

Controlliamo che questa colorazione soddisfi le ipotesi del lemma di Sperner. Innanzitutto, il vertice $\boldsymbol{e}_i$ deve ricevere come colore i, poiché la sola componente negativa possibile di $f(\boldsymbol{e}_i) - \boldsymbol{e}_i$ è la componente i-esima. Inoltre, se $\boldsymbol{v}$ giace sullo spigolo opposto a $\boldsymbol{e}_i$, allora $v_i = 0$, cosicché la componente i-esima di $f(\boldsymbol{v}) - \boldsymbol{v}$ non può essere negativa e dunque $\boldsymbol{v}$ non ottiene il colore i.

Il lemma di Sperner ci dice che in ogni triangolazione $\mathcal{T}_k$ vi è un triangolo tricolore $\{\boldsymbol{v}^{k:1}, \boldsymbol{v}^{k:2}, \boldsymbol{v}^{k:3}\}$ con $\lambda(\boldsymbol{v}^{k:i}) = i$. La successione dei punti $(\boldsymbol{v}^{k:1})_{k\geq 1}$ non deve convergere, ma dal momento che il simplesso Δ è compatto qualche sottosuccessione ha un punto limite. Dopo aver sostituito la successione delle triangolazioni $\mathcal{T}_k$ con la sottosuccessione corrispondente (che per semplicità indichiamo anche con $\mathcal{T}_k$) possiamo supporre che $(\boldsymbol{v}^{k:1})_k$ converga ad un punto $\boldsymbol{v} \in \Delta$. Ora la distanza di $\boldsymbol{v}^{k:2}$ e $\boldsymbol{v}^{k:3}$ da $\boldsymbol{v}^{k:1}$ è al massimo l'ampiezza $\delta(\mathcal{T}_k)$ della griglia, che converge a 0. Pertanto le successioni $(\boldsymbol{v}^{k:2})$ e $(\boldsymbol{v}^{k:3})$ convergono allo *stesso* punto $\boldsymbol{v}$.

Ma dov'è $f(\boldsymbol{v})$? Sappiamo che la prima coordinata $f(\boldsymbol{v}^{k:1})$ è minore di quella di $\boldsymbol{v}^{k:1}$ per ogni k. Ora poiché f è continua, deriviamo che la prima coordinata di $f(\boldsymbol{v})$ è minore o uguale a quella di $\boldsymbol{v}$. Lo stesso ragionamento funziona per la seconda e la terza coordinata. Pertanto nessuna delle coordinate di $f(\boldsymbol{v}) - \boldsymbol{v}$ è positiva — ed abbiamo già visto che ciò contraddice l'ipotesi $f(\boldsymbol{v}) \neq \boldsymbol{v}$. □

Bibliografia

[1] L. E. J. BROUWER: *Über Abbildungen von Mannigfaltigkeiten,* Math. Annalen **71** (1912), 97-115.

[2] W. G. BROWN: *On graphs that do not contain a Thomsen graph,* Canadian Math. Bull. **9** (1966), 281-285.

[3] P. ERDŐS, A. RÉNYI & V. SÓS: *On a problem of graph theory,* Studia Sci. Math. Hungar. **1** (1966), 215-235.

[4] P. ERDŐS & G. SZEKERES: *A combinatorial problem in geometry,* Compositio Math. (1935), 463-470.

[5] S. HOŞTEN & W. D. MORRIS: *The order dimension of the complete graph,* Discrete Math. **201** (1999), 133-139.

[6] I. REIMAN: *Über ein Problem von K. Zarankiewicz,* Acta Math. Acad. Sci. Hungar. **9** (1958), 269-273.

[7] J. SPENCER: *Minimal scrambling sets of simple orders,* Acta Math. Acad. Sci. Hungar. **22** (1971), 349-353.

[8] E. SPERNER: *Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes,* Abh. Math. Sem. Hamburg **6** (1928), 265-272.

[9] W. T. TROTTER: *Combinatorics and Partially Ordered Sets: Dimension Theory,* John Hopkins University Press, Baltimore and London 1992.

Tre celebri teoremi sugli insiemi finiti

Capitolo 23

In questo capitolo trattiamo un tema basilare del calcolo combinatorio: le proprietà e le dimensioni di famiglie speciali $\mathcal{F}$ di sottoinsiemi di un insieme finito $N = \{1, 2, \ldots, n\}$. Cominciamo con due risultati classici in questo campo: i teoremi di Sperner e di Erdős-Ko-Rado; essi hanno in comune il fatto che furono ridimostrati varie volte e che ognuno di loro inaugurò un nuovo campo della teoria combinatoria degli insiemi. Per entrambi i teoremi, l'induzione sembra essere il metodo naturale, ma gli argomenti che discuteremo sono ben diversi e davvero brillanti.

Emanuel Sperner

Nel 1928 Emanuel Sperner pose la domanda seguente (e vi rispose): supponiamo di avere l'insieme $N = \{1, 2, \ldots, n\}$. Chiamiamo *anticatena* una famiglia $\mathcal{F}$ di sottoinsiemi di N tale che nessun insieme di $\mathcal{F}$ contiene un altro insieme della famiglia $\mathcal{F}$. Qual è la dimensione massima di un'anticatena? Chiaramente, la famiglia $\mathcal{F}_k$ di tutti i k-insiemi soddisfa la proprietà di anticatena con $|\mathcal{F}_k| = \binom{n}{k}$. Considerando il più grande dei coefficienti binomiali (si veda a pagina 12) concludiamo che vi è un'anticatena di dimensione $\binom{n}{\lfloor n/2 \rfloor} = \max_k \binom{n}{k}$. Il teorema di Sperner afferma che non ve ne sono di più grandi.

Teorema 1. *La dimensione massima di un'anticatena di un n-insieme è $\binom{n}{\lfloor n/2 \rfloor}$.*

■ **Dimostrazione.** Fra le varie dimostrazioni la seguente, dovuta a David Lubell, è probabilmente la più breve ed elegante. Sia $\mathcal{F}$ un'anticatena arbitraria. Allora dobbiamo dimostrare che $|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}$. La chiave della dimostrazione è considerare *catene* di sottoinsiemi $\varnothing = C_0 \subset C_1 \subset C_2 \subset \ldots \subset C_n = N$, dove $|C_i| = i$ per $i = 0, \ldots, n$. Quante catene ci sono? Chiaramente, otteniamo una catena sommando uno a uno gli elementi di N, pertanto vi sono esattamente tante catene quante permutazioni di N, ovvero $n!$. Ora, per un insieme $A \in \mathcal{F}$, determiniamo quante di queste catene contengano A. Anche questo è semplice. Per passare da $\varnothing$ ad A dobbiamo aggiungere gli elementi di A uno ad uno e quindi per passare da A ad N dobbiamo aggiungere gli elementi rimanenti. Pertanto se A contiene k elementi, allora considerando tutte queste coppie di catene collegate fra loro vediamo che vi sono esattamente $k!(n-k)!$ catene di questo tipo. Si noti che nessuna catena può attraversare due insiemi diversi A e B di $\mathcal{F}$, poiché $\mathcal{F}$ è un'anticatena.

Per completare la dimostrazione, sia m_k il numero di k-insiemi in $\mathcal{F}$. Allora $|\mathcal{F}| = \sum_{k=0}^{n} m_k$. Dalla nostra discussione segue che il numero di catene

che attraversano un membro di $\mathcal{F}$ è

$$\sum_{k=0}^{n} m_k \, k! \, (n-k)!,$$

e tale espressione non può superare il numero $n!$ di *tutte* le catene. Pertanto,

$$\sum_{k=0}^{n} m_k \frac{k!(n-k)!}{n!} \;\leq\; 1, \qquad \text{oppure} \qquad \sum_{k=0}^{n} \frac{m_k}{\binom{n}{k}} \;\leq\; 1.$$

Sostituendo i denominatori con il più grande dei coefficienti binomiali otteniamo quindi

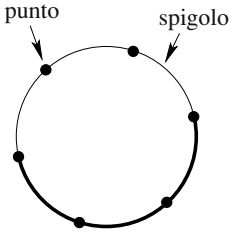
$$\frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \sum_{k=0}^{n} m_k \;\leq\; 1, \qquad \text{ovvero} \qquad |\mathcal{F}| \;=\; \sum_{k=0}^{n} m_k \;\leq\; \binom{n}{\lfloor n/2 \rfloor},$$

e la dimostrazione è completa. □

Si verifichi che la famiglia di tutti gli $\frac{n}{2}$-insiemi per n pari e, rispettivamente, le due famiglie di tutti gli $\frac{n-1}{2}$-insiemi e di tutti gli $\frac{n+1}{2}$-insiemi, quando n è dispari, sono effettivamente le *sole* anticatene di dimensione massima!

Il nostro secondo risultato è di natura completamente diversa. Consideriamo nuovamente l'insieme $N = \{1, \ldots, n\}$. Diciamo che una famiglia $\mathcal{F}$ di sottoinsiemi è una *famiglia intersecante* se ogni coppia di insiemi di $\mathcal{F}$ ha almeno un elemento in comune. È quasi immediato notare che la dimensione massima di una famiglia intersecante è 2^{n-1}. Se $A \in \mathcal{F}$, allora il complemento $A^c = N \backslash A$ ha intersezione vuota con A e dunque non può essere in $\mathcal{F}$. Pertanto concludiamo che una famiglia intersecante contiene al massimo la metà del numero 2^n di tutti i sottoinsiemi, ovvero, $|\mathcal{F}| \leq 2^{n-1}$. D'altro canto, se consideriamo la famiglia di tutti gli insiemi contenenti un elemento fisso, diciamo la famiglia $\mathcal{F}_1$ di tutti gli insiemi contenenti 1, allora chiaramente $|\mathcal{F}_1| = 2^{n-1}$ ed il problema è risolto.

Poniamoci ora la seguente domanda: quanto può essere grande una famiglia intersecante $\mathcal{F}$ se tutti gli insiemi in $\mathcal{F}$ hanno la stessa dimensione, diciamo k? Chiameremo tali famiglie *k-famiglie intersecanti*. Per evitare casi banali, supponiamo $n \geq 2k$ poiché altrimenti qualunque coppia di k-insiemi si intersecherebbe e non vi sarebbe nulla da dimostrare. Riprendendo l'idea esposta sopra, otteniamo certamente una tale famiglia, $\mathcal{F}_1$, considerando tutti i k-insiemi contenenti un elemento fissato, ad esempio 1. Chiaramente, otteniamo tutti gli insiemi in $\mathcal{F}_1$ sommando a 1 tutti i $(k-1)$-sottoinsiemi di $\{2, 3, \ldots, n\}$, pertanto $|\mathcal{F}_1| = \binom{n-1}{k-1}$. Possiamo fare di meglio? Il teorema di Erdős-Ko-Rado ci dice di no.



Un cerchio C per $n = 6$. I lati in grassetto indicano un arco di lunghezza 3

Teorema 2. *La dimensione massima di una k-famiglia intersecante in un n-insieme è $\binom{n-1}{k-1}$ quando $n \geq 2k$.*

Paul Erdős, Chao Ko e Richard Rado trovarono questo risultato nel 1938, ma esso non fu pubblicato che 23 anni più tardi. Da allora è stata data una moltitudine di dimostrazioni e varianti, ma l'argomento che segue, dovuto a Gyula Katona, è particolarmente elegante.

■ **Dimostrazione.** La chiave alla dimostrazione è il semplice lemma seguente, che in un primo momento sembra totalmente scollegato dal nostro problema. Si consideri un cerchio C diviso da n punti in n lati. Chiamiamo *arco* di lunghezza k l'unione fra $k+1$ punti consecutivi e i k lati situati fra loro.

Lemma. *Sia $n \geq 2k$, e si supponga di avere t archi distinti $A_1, \ldots, A_t$ di lunghezza k, tali che ogni coppia di archi abbia un lato in comune. Allora $t \leq k$.*

Per dimostrare il lemma, si noti dapprima che ogni punto di C è l'estremità di non più di un arco. In effetti, se A_i e A_j avessero un'estremità v in comune, allora dovrebbero puntare verso direzioni diverse (poiché sono distinti). Ma in tal caso non potrebbero avere un lato in comune poiché $n \geq 2k$. Fissiamo A_1. Dal momento che ogni A_i $(i \geq 2)$ ha un lato in comune con A_1, una delle estremità di A_i è un punto interno di A_1. Poiché queste estremità devono essere distinte come abbiamo appena visto, e poiché A_1 contiene $k-1$ punti interni, concludiamo che possono esistere al massimo $k-1$ altri archi e pertanto complessivamente al massimo k archi. □

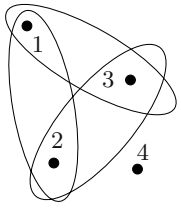
Continuamo ora con la dimostrazione del teorema di Erdős-Ko-Rado. Sia $\mathcal{F}$ una k-famiglia intersecante. Si consideri un cerchio C con n punti e n lati come sopra. Prendiamo un ciclo qualunque $\pi = (a_1, a_2, \ldots, a_n)$ e scriviamo i numeri a_i in senso orario accanto ai lati di C. Contiamo il numero di insiemi $A \in \mathcal{F}$ che appaiono come k numeri *consecutivi* su C. Poiché $\mathcal{F}$ è una famiglia intersecante, grazie al nostro lemma otteniamo al massimo k di tali insiemi. Poiché questo vale per ogni ciclo, e poiché esistono $(n-1)!$ cicli, produciamo in questo modo al più

$$k(n-1)!$$

insiemi di $\mathcal{F}$ che appaiono come elementi consecutivi di un ciclo. Quanto spesso contiamo un insieme fissato $A \in \mathcal{F}$? È facile: A appare in π se i k elementi di A appaiono consecutivamente in un ordine dato. Pertanto abbiamo $k!$ possibilità di scrivere A consecutivamente, e $(n-k)!$ modi di ordinare gli elementi restanti. Concludiamo dunque che un insieme fissato A appare in esattamente $k!(n-k)!$ cicli, e dunque che

$$|\mathcal{F}| \leq \frac{k(n-1)!}{k!(n-k)!} = \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} = \binom{n-1}{k-1}. \quad \square$$

Di nuovo possiamo chiederci se le famiglie contenenti un elemento fisso siano le sole k-famiglie intersecanti di dimensione massima. Ciò è certamente falso per $n = 2k$. Per esempio, per $n = 4$ e $k = 2$ la famiglia $\{1,2\},\{1,3\},\{2,3\}$ ha anch'essa dimensione $\binom{3}{1} = 3$. Più in generale, per $n = 2k$ le più grandi k-famiglie intersecanti hanno dimensione $\frac{1}{2}\binom{n}{k} = \binom{n-1}{k-1}$, e sono composte nel modo seguente: si considerano tutte le coppie formate da un k-insieme A e dal suo complementare $N \backslash A$. Ma per $n > 2k$ le famiglie contenenti un elemento fisso sono effettivamente le uniche. Il lettore è invitato a cimentarsi con la dimostrazione.



Una famiglia intersecante per $n = 4$, $k = 2$

"Un matrimonio collettivo"

In conclusione, ci dedichiamo al terzo risultato che è incontestabilmente il più importante teorema di base nella teoria degli insiemi finiti, il "teorema delle nozze" che Philip Hall dimostrò nel 1935. Esso aprì la porta a quella che oggi è definita la teoria dell'accoppiamento, con una gran varietà di applicazioni; ne vedremo alcune via via che procediamo.

Si consideri un insieme finito X ed una collezione $A_1, \ldots, A_n$ di sottoinsiemi di X (non necessariamente distinti). Una successione $x_1, \ldots, x_n$ si definisce un *sistema di rappresentanti distinti* di $\{A_1, \ldots, A_n\}$ se gli x_i sono elementi distinti di X, e se $x_i \in A_i$ per ogni i. Ovviamente, un tale sistema, abbreviato con SRD, potrebbe non esistere, per esempio quando uno degli insiemi A_i è vuoto. Il teorema di Hall fornisce la condizione precisa per l'esistenza di un SRD.

Prima di dare il risultato esponiamo l'interpretazione "umana" che gli fece attribuire il nome folcloristico di *teorema delle nozze*: si consideri un insieme $\{1, \ldots, n\}$ di ragazze ed un insieme X di ragazzi. Ogniqualvolta $x \in A_i$, allora la ragazza i ed il ragazzo x sono inclini a sposarsi, pertanto A_i è semplicemente l'insieme dei fidanzamenti possibili della ragazza i. Un SRD rappresenta allora un matrimonio collettivo in cui ogni ragazza sposa un ragazzo che le piace.

Tornando agli insiemi, ecco l'enunciato del risultato.

Teorema 3. *Sia $A_1, \ldots, A_n$ una collezione di sottoinsiemi di un insieme finito X. Allora esiste un insieme di rappresentanti distinti se e solo se l'unione di m insiemi A_i presi in modo qualunque contiene almeno m elementi, con $1 \leq m \leq n$.*

La condizione è chiaramente necessaria: se m insiemi A_i contengono fra loro meno di m elementi, allora questi m insiemi non possono certo essere rappresentati da elementi distinti. Il fatto sorprendente (che conferisce al teorema un'applicabilità universale) è che questa condizione ovvia è anche sufficiente. La dimostrazione originale di Hall era piuttosto complessa; successivamente furono date molte altre dimostrazioni. La più naturale è probabilmente la seguente (dovuta a Easterfield e riscoperta da Halmos e Vaughan).

■ **Dimostrazione.** Usiamo l'induzione su n. Per $n = 1$ non vi è nulla da dimostrare. Sia $n > 1$ e supponiamo che $\{A_1, \ldots, A_n\}$ verifichi la condizione del teorema (che abbrevieremo con (H) nel seguito). Una collezione di ℓ insiemi A_i con $1 \leq \ell < n$ è detta una *famiglia critica* se la sua unione ha cardinalità ℓ. Si possono presentare due casi.

Caso 1: Non vi è alcuna famiglia critica.

Si scelga un elemento qualsiasi $x \in A_n$. Si rimuova x da X e si consideri la collezione $A'_1, \ldots, A'_{n-1}$ con $A'_i = A_i \backslash \{x\}$. Poiché non vi è alcuna famiglia critica, l'unione di qualunque gruppo di m insiemi A'_i contiene almeno m elementi. Pertanto per induzione su n dimostriamo che esiste un SRD $x_1, \ldots, x_{n-1}$ di $\{A'_1, \ldots, A'_{n-1}\}$. Aggiungendovi $x_n = x$, ciò fornisce un SRD per la collezione originale.

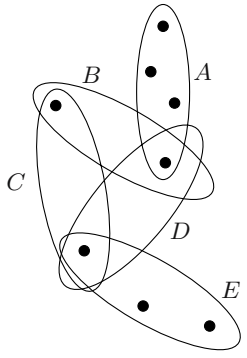
Caso 2: Esiste almeno una famiglia critica.

Dopo aver eventualmente rinumerato gli insiemi, possiamo supporre che $\{A_1, \ldots, A_\ell\}$ è una famiglia critica. Allora $\bigcup_{i=1}^{\ell} A_i = \widetilde{X}$ con $|\widetilde{X}| = \ell$. Poiché $\ell < n$, deduciamo per induzione l'esistenza di un SRD per $A_1, \ldots, A_\ell$, ovvero, equivalentemente, abbiamo trovato una numerazione $x_1, \ldots, x_\ell$ di $\widetilde{X}$ tale che $x_i \in A_i$ per ogni $i \leq \ell$.

Si consideri ora la collezione restante $A_{\ell+1}, \ldots, A_n$, e si prendano (in modo arbitrario) m di tali insiemi. Poiché l'unione di $A_1, \ldots, A_\ell$ e di questi m insiemi contiene almeno $\ell + m$ elementi per via della condizione (H), deduciamo che gli m insiemi contengono almeno m elementi al di fuori di $\widetilde{X}$. In altre parole, la condizione (H) è soddisfatta per la famiglia

$$A_{\ell+1}\backslash\widetilde{X},\ \ldots\ , A_n\backslash\widetilde{X}.$$

L'induzione dà ora un SRD per $A_{\ell+1}, \ldots, A_n$ che evita $\widetilde{X}$. Combinandolo con $x_1, \ldots, x_\ell$ otteniamo un SRD per ogni insieme A_i. Questo completa la dimostrazione. □



$\{B, C, D\}$ è una famiglia critica

Come già detto, il teorema di Hall diede origine al dominio (ora assai vasto) della teoria dell'accoppiamento [6]. Fra le molte varianti e ramificazioni riportiamo un risultato particolarmente interessante che il lettore è invitato a dimostrare da sé:

> *Si supponga che gli insiemi $A_1, \ldots, A_n$ abbiano tutti cardinalità $k \geq 1$ ed inoltre che nessun elemento sia contenuto in più di k insiemi. Allora esistono k SRD tali che per ogni i, i k rappresentanti di A_i siano distinti e pertanto formino l'insieme A_i.*

Si tratta di uno splendido risultato che dovrebbe aprire nuovi orizzonti sulle possibilità di matrimonio.

Bibliografia

[1] T. E. EASTERFIELD: *A combinatorial algorithm,* J. London Math. Soc. **21** (1946), 219-226.

[2] P. ERDŐS, C. KO & R. RADO: *Intersection theorems for systems of finite sets,* Quart. J. Math. (Oxford), Ser. (2) **12** (1961), 313-320.

[3] P. HALL: *On representatives of subsets,* J. London Math. Soc. **10** (1935), 26-30.

[4] P. R. HALMOS & H. E. VAUGHAN: *The marriage problem,* Amer. J. Math. **72** (1950), 214-215.

[5] G. KATONA: *A simple proof of the Erdős-Ko-Rado theorem,* J. Combinatorial Theory, Ser. B **13** (1972), 183-184.

[6] L. LOVÁSZ & M. D. PLUMMER: *Matching Theory,* Akadémiai Kiadó, Budapest 1986.

[7] D. LUBELL: *A short proof of Sperner's theorem,* J. Combinatorial Theory **1** (1966), 299.

[8] E. SPERNER: *Ein Satz über Untermengen einer endlichen Menge,* Math. Zeitschrift **27** (1928), 544-548.

Mescolare le carte

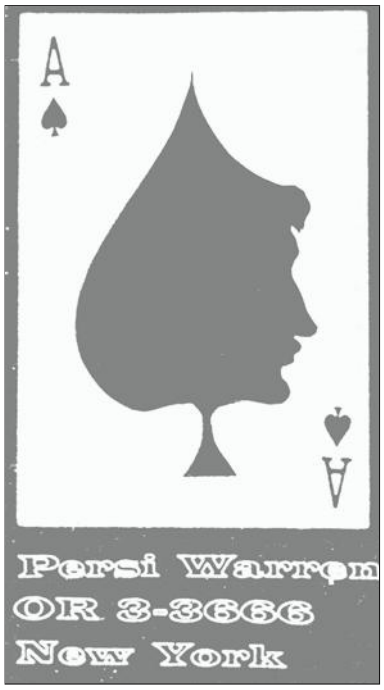
Capitolo 24

> *Quante volte bisogna mischiare un mazzo di carte perché esso sia "a caso" ?*

L'analisi di processi aleatori è una situazione piuttosto comune nella vita ("Quanto ci vuole per arrivare in aeroporto durante l'ora di punta?") come nella matematica. Ovviamente, per ottenere risposte sensate a tali problemi si dovranno formulare domande in modo sensato. Per il problema del mescolamento delle carte, ciò significa che dobbiamo

- specificare le dimensioni del mazzo (ad esempio, $n = 52$ carte),
- definire il modo in cui mescoliamo (analizzeremo prima mescolamenti top-in-at-random e poi i più realistici ed efficaci mescolamenti "all'americana" (NdT: "riffle-shuffle" nell'edizione inglese) ed infine
- spiegare cosa intendiamo con "a caso" o "quasi a caso".

Il nostro obiettivo in questo capitolo è un'analisi del riffle shuffle, dovuta a Edgar N. Gilbert e Claude Shannon (1955, non pubblicato) e Jim Reeds (1981, non pubblicato), seguendo lo statistico David Aldous e l'ex prestigiatore divenuto matematico Persi Diaconis, secondo [1]. Non raggiungeremo il risultato finale esatto che 7 mescolamenti all'americana *sono* sufficienti per ottenere un mazzo di $n = 52$ molto prossimo ad essere aleatorio, mentre 6 mescolamenti all'americana non bastano — ma otterremo un limite superiore di 12 e vedremo strada facendo alcune idee estremamente belle: i concetti di regole d'arresto e di "tempo forte uniforme", il lemma che il tempo forte uniforme maggiora la distanza in variazione, il lemma di inversione di Reeds e dunque l'interpretazione di mescolamento come "riordinamento al contrario". Alla fine, tutto si ridurrà a due problemi combinatori molto classici, ovvero quello del collezionista di figurine ed il paradosso dei compleanni. Cominciamo quindi da questi!



Il biglietto da visita di Persi Diaconis da mago. In un'intervista successiva affermò: "Se dite che siete un professore a Stanford la gente vi tratta con rispetto. Se dite che inventate giochi di prestigio, non vi presentano nemmeno a loro figlia"

Il paradosso del compleanno

Si scelgano n persone in modo arbitrario — i partecipanti ad un corso o a un seminario, ad esempio. Qual è la probabilità che esse compiano gli anni

tutte in giorni diversi? Con le solite ipotesi di semplificazione (365 giorni in un anno, nessun effetto stagionale, niente gemelli) la probabilità è

$$p(n) \;=\; \prod_{i=1}^{n-1}\Big(1-\frac{i}{365}\Big),$$

che è inferiore a $\frac{1}{2}$ per $n = 23$ (questo è il "paradosso del compleanno"!) minore del 9 percento per $n = 42$, ed esattamente 0 per $n > 365$ (il "principio del casellario", si veda al Capitolo 22). La formula è semplice da verificare, se prendiamo le persone in un ordine dato: se le prime i persone hanno compleanni diversi, allora la persona $(i+1)$-esima che non rovina la serie è $1-\frac{i}{365}$, poiché rimangono $365-i$ compleanni.

Analogamente, se n palline sono distribuite indipendentemente ed arbitrariamente in K scatole, allora la probabilità che nessuna scatola contenga più di una pallina è

$$p(n,K) \;=\; \prod_{i=1}^{n-1}\Big(1-\frac{i}{K}\Big).$$

Il collezionista di figurine

I bambini comprano figurine delle star (o dei calciatori) per i loro album in bustine non trasparenti, così non sanno che figurine avranno. Se ci sono n figurine diverse, qual è il numero atteso di figurine che un ragazzino deve comprare per avere almeno un esemplare di ogni figurina?

Equivalentemente, se estraete arbitrariamente delle palline da un recipiente che contiene n palline distinguibili e se reinserite la pallina ogni volta e quindi rimescolate bene, quante estrazioni dovete effettuare in media per estrarre ogni pallina almeno una volta?

Se avete già estratto k palline diverse, allora la probabilità di non averne una nuova nella prossima estrazione è $\frac{k}{n}$. Pertanto la probabilità di aver bisogno di esattamente s estrazioni per la prossima nuova pallina è $(\frac{k}{n})^{s-1}(1-\frac{k}{n})$. Dunque il numero atteso di estrazioni per la prossima pallina nuova è

$$\sum_{s\ge1}\Big(\frac{k}{n}\Big)^{s-1}\Big(1-\frac{k}{n}\Big)s \;=\; \frac{1}{1-\frac{k}{n}},$$

come deduciamo dalla serie a margine. Quindi il numero atteso di estrazioni per aver estratto *ognuna* delle diverse palline almeno una volta è

$$\sum_{k=0}^{n-1}\frac{1}{1-\frac{k}{n}} \;=\; \frac{n}{n}+\frac{n}{n-1}+\cdots+\frac{n}{2}+\frac{n}{1} \;=\; nH_n \;\approx\; n\log n,$$

grazie alle maggiorazioni sui numeri armonici che abbiamo ottenuto a pagina 11. Pertanto la risposta al problema del collezionista di figurine è che dobbiamo aspettarci che siano necessarie circa $n\log n$ estrazioni.

$$\begin{aligned}\sum_{s\ge1}x^{s-1}(1-x)s \;&=\\ &=\; \sum_{s\ge1}x^{s-1}s \;-\; \sum_{s\ge1}x^{s}s\\ &=\; \sum_{s\ge0}x^{s}(s+1) \;-\; \sum_{s\ge0}x^{s}s\\ &=\; \sum_{s\ge0}x^{s} \;=\; \frac{1}{1-x},\end{aligned}$$

dove alla fine otteniamo una serie geometrica (si veda a pagina 34)

La stima di cui avremo bisogno è per la probabilità che servano un numero significativamente maggiore di $n \log n$ tentativi. Se V_n indica il numero di estrazioni necessarie (questa è la variabile aleatoria il cui valore atteso è $E[V_n] \approx n \log n$), allora per $n \geq 1$ e $c \geq 0$ la probabilità che ci servano più di $m := \lceil n \log n + cn \rceil$ estrazioni è

$$\text{Prob}\big[V_n > m\big] \;\leq\; e^{-c}.$$

In effetti, se A_i indica l'evento in cui la pallina i non sia estratta nelle prime m estrazioni, allora

$$\begin{aligned} \text{Prob}\big[V_n > m\big] \;&=\; \text{Prob}\Big[\bigcup_i A_i\Big] \;\leq\; \sum_i \text{Prob}\big[A_i\big] \\ &=\; n\Big(1 - \frac{1}{n}\Big)^m \;<\; ne^{-m/n} \;\leq\; e^{-c}. \end{aligned}$$

Un po' di analisi mostra che $\big(1 - \frac{1}{n}\big)^n$ è una funzione crescente in n, che converge a $1/e$. Pertanto $\big(1 - \frac{1}{n}\big)^n \;<\; \frac{1}{e}$ per ogni $n \geq 1$

Ora prendiamo in mano un mazzo di n carte. Le numeriamo da 1 a n nell'ordine in cui si presentano, pertanto la carta numerata "1" è in cima al mazzo, mentre "n" è in fondo. Indichiamo d'ora in poi con $\mathfrak{S}_n$ l'insieme di tutte le permutazioni di $1, \ldots, n$. *Mescolare* il mazzo equivale ad effettuare certe *permutazioni arbitrarie* all'ordine delle carte. Idealmente, questo potrebbe significare che applichiamo una permutazione arbitraria $\pi \in \mathfrak{S}_n$ al nostro ordine iniziale $(1, 2, \ldots, n)$, ognuna delle quali con la stessa probabilità $\frac{1}{n!}$. Pertanto, dopo averlo fatto una sola volta, avremmo il nostro mazzo nell'ordine $\pi = (\pi(1), \pi(2), \ldots, \pi(n))$ e questo sarebbe un perfetto ordine aleatorio. Ma non è ciò che succede nella realtà. Piuttosto, quando si mescola si verificano solo "certe" permutazioni, forse non tutte con la stessa probabilità, e questo si ripete un "certo" numero di volte. Dopo di che, ci aspettiamo o speriamo che il mazzo sia almeno "vicino all'essere aleatorio".

Mescolamenti top-in-at-random

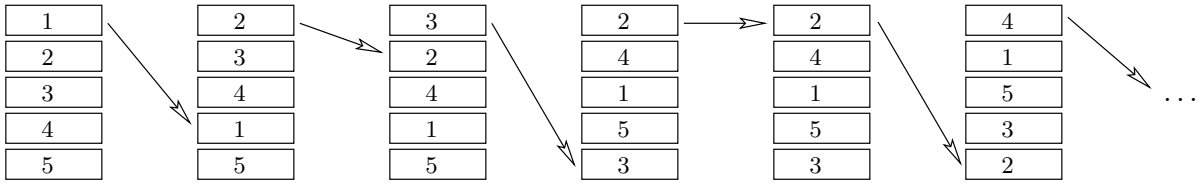
Mescolamento top-in-at-random

Questo si effettua come segue: prendete la prima carta in cima al mazzo e inseritela nel mazzo in uno degli n diversi punti possibili, ognuno di essi avente probabilità $\frac{1}{n}$. Si è pertanto applicata una delle permutazioni

$$\tau_i \;=\; \big(2, 3, \ldots, \overset{\overset{i}{\downarrow}}{i}, 1, i{+}1, \ldots, n\big),$$

(per $1 \leq i \leq n$). Dopo uno di questi mescolamenti il mazzo non appare in ordine aleatorio ed in effetti ci aspettiamo di aver bisogno di molti mescolamenti del genere per raggiungere quell'obiettivo.

Un tipico giro di mescolamenti top-in-at-random può apparire nel modo seguente (per $n = 5$):



Come dovremmo misurare quanto si è "vicino all'essere aleatorio"? I probabilisti hanno coniato la "distanza in variazione" come misura alquanto rigorosa dell'aleatorietà: consideriamo la distribuzione di probabilità sugli $n!$ diversi ordinamenti possibili del nostro mazzo o, equivalentemente, sulle $n!$ diverse permutazioni $\sigma \in \mathfrak{S}_n$ che danno gli ordinamenti.

Due esempi sono la nostra distribuzione iniziale E, data da

$$\begin{aligned} \mathsf{E}(\mathrm{id}) &= 1, \\ \mathsf{E}(\pi) &= 0 \quad \text{altrimenti,} \end{aligned}$$

e la distribuzione uniforme U data da

$$\mathsf{U}(\pi) = \tfrac{1}{n!} \quad \text{per ogni } \pi \in \mathfrak{S}_n.$$

La *distanza in variazione* fra due distribuzioni di probabilità Q_1 e Q_2 è definita ora come

$$\|\mathsf{Q}_1 - \mathsf{Q}_2\| := \tfrac{1}{2} \sum_{\pi \in \mathfrak{S}_n} |\mathsf{Q}_1(\pi) - \mathsf{Q}_2(\pi)|.$$

Ponendo $S := \{\pi \in \mathfrak{S}_n : \mathsf{Q}_1(\pi) > \mathsf{Q}_2(\pi)\}$ ed usando $\sum_\pi \mathsf{Q}_1(\pi) = \sum_\pi \mathsf{Q}_2(\pi) = 1$ si ottiene

$$\|\mathsf{Q}_1 - \mathsf{Q}_2\| = \max_{S \subseteq \mathfrak{S}_n} |\mathsf{Q}_1(S) - \mathsf{Q}_2(S)|,$$

con $\mathsf{Q}_i(S) := \sum_{\pi \in S} \mathsf{Q}_i(\pi)$. Abbiamo chiaramente $0 \le \|\mathsf{Q}_1 - \mathsf{Q}_2\| \le 1$. Nel seguito, "vicino all'essere aleatorio" sarà interpretato come "avente una piccola distanza della variazione dalla distribuzione uniforme". Qui la distanza tra la distribuzione iniziale e la distribuzione uniforme è molto vicina a 1:

$$\|\mathsf{E} - \mathsf{U}\| = 1 - \tfrac{1}{n!}.$$

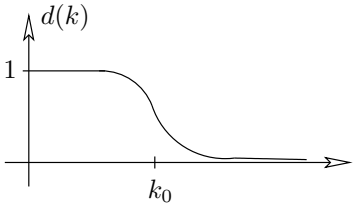
Dopo un mescolamento top-in-at-random, questo valore non sarà molto migliore:

$$\|\mathsf{Top} - \mathsf{U}\| = 1 - \tfrac{1}{(n-1)!}.$$

La distribuzione di probabilità su $\mathfrak{S}_n$ che otteniamo applicando il mescolamento top-in-at-random k volte sarà indicata con Top^{*k}. Come si comporta quindi $\|\mathsf{Top}^{*k} - \mathsf{U}\|$ se k diventa più grande, ovvero se ripetiamo il mescolamento? E per altri tipi di mescolamenti? La teoria generale (in particolare le catene di Markov sui gruppi finiti su cui si veda ad esempio Behrends [3]) implica che per k grandi la distanza in variazione $d(k) := \|\mathsf{Top}^{*k} - \mathsf{U}\|$ tende a zero con velocità esponenziale, ma non dà il fenomeno di "cut-off" che si osserva nella pratica: dopo un certo numero k_0 di mescolamenti, "improvvisamente" $d(k)$ tende a zero piuttosto velocemente. A margine mostriamo uno schizzo schematico della situazione.

Per i giocatori, la domanda non è "esattamente quanto è vicino all'uniformità il mazzo dopo un milione di mescolamenti all'americana?", ma "bastano 7 mescolamenti?"

(Aldous & Diaconis [1])



Regole d'arresto forti uniformi

L'idea stupefacente di regole d'arresto forti uniformi di Aldous e Diaconis cattura le caratteristiche essenziali. Immaginiamo che il manager del casinò controlli da vicino i processi di mescolamento, analizzi le specifiche permutazioni applicate al mazzo ad ogni passo e dopo un numero di passi che dipende dalle permutazioni che ha visto dica "STOP!". Quindi egli ha una *regola d'arresto* che termina il processo di mescolamento. Essa dipende solo dai mescolamenti (arbitrari) che sono stati già effettuati. La regola d'arresto è *forte uniforme* se la seguente condizione vale per ogni $k \geq 0$:

> ***Se** il processo è arrestato dopo esattamente k passi, **allora** le permutazioni risultanti del mazzo hanno distribuzione uniforme (esattamente!).*

Sia T il numero di passi effettuati finché la regola d'arresto dica al manager di gridare "STOP!"; questa è dunque una variabile aleatoria. Analogamente, l'ordine del mazzo dopo k mescolamenti è dato da una variabile aleatoria X_k (con valori in $\mathfrak{S}_n$). Con questo, la regola d'arresto è forte uniforme se per ogni valore fattibile di k,

$$\text{Prob}\big[X_k = \pi \mid T = k\big] \;=\; \frac{1}{n!} \quad \text{per ogni } \pi \in \mathfrak{S}_n.$$

Tre aspetti rendono questa definizione interessante, utile, e notevole:

1. Esistono regole d'arresto forti uniformi: per molti esempi sono alquanto semplici.
2. Esse possono essere analizzate: il tentativo di determinare $\text{Prob}[T > k]$ porta spesso a semplici problemi combinatori.
3. Fornisce maggiorazioni efficaci per le distanze in variazione quali ad esempio $d(k) = \|\mathsf{Top}^{*k} - \mathsf{U}\|$.

Probabilità condizionale

La *probabilità condizionale*

$$\text{Prob}[A \mid B]$$

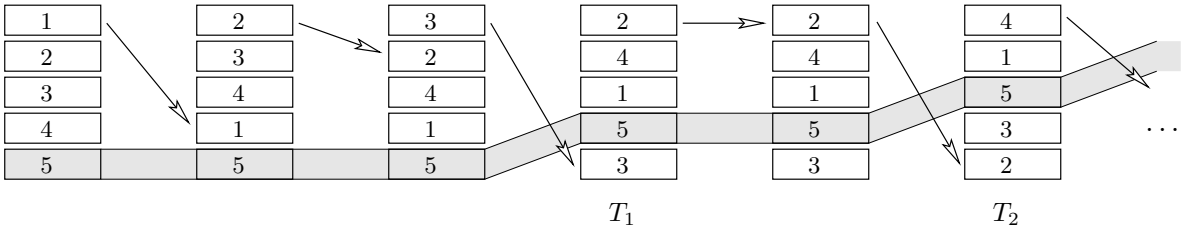
indica la probabilità dell'evento A sotto la condizione che accada B. Questa è semplicemente la probabilità che entrambi gli eventi si verifichino, divisa per la probabilità che B sia vero, ovvero,

$$\text{Prob}[A \mid B] \;=\; \frac{\text{Prob}[A \wedge B]}{\text{Prob}[B]}.$$

Per esempio, per i mescolamenti top-in-at-random una regola d'arresto forte uniforme è:

> "STOP dopo che la carta inizialmente in fondo (denominata n) è riinserita per la prima volta nel mazzo."

In effetti, se rintracciamo la carta n durante questi mescolamenti,



vediamo che durante l'intero processo l'ordine delle carte al di sotto di questa carta è completamente uniforme. Pertanto, dopo che la carta n è risalita

in cima e quindi è inserita aleatoriamente, il mazzo è distribuito uniformemente; semplicemente non sappiamo quando esattamente ciò succeda (ma il manager lo sa).

Ora sia T_i la variabile aleatoria che conta il numero di mescolamenti effettuati finché per la prima volta i carte si trovino sotto la carta n. Dobbiamo quindi determinare la distribuzione di

$$T \;=\; T_1 + (T_2 - T_1) + \ldots + (T_{n-1} - T_{n-2}) + (T - T_{n-1}).$$

Ma ogni addendo in questa somma corrisponde ad un problema del collezionista di figurine: $T_i - T_{i-1}$ è il tempo necessario affinché la carta in cima sia inserita in una delle i posizioni possibili sotto la carta n. Questo è anche il tempo necessario perché il collezionista di figurine passi dalla figurina $(n-i)$-esima a quella $(n-i+1)$-esima. Sia V_i il numero di figurine acquistate prima di avere i figurine diverse. Allora

$$V_n \;=\; V_1 + (V_2 - V_1) + \ldots + (V_{n-1} - V_{n-2}) + (V_n - V_{n-1}),$$

ed abbiamo visto che $\text{Prob}[T_i - T_{i-1} = j] = \text{Prob}[V_{n-i+1} - V_{n-i} = j]$ per ogni i e j. Pertanto il collezionista ed il mescolatore top-in-at-random realizzano successioni equivalenti di processi aleatori indipendenti, ma in ordine inverso (per il collezionista, il difficile è alla fine). Pertanto sappiamo che la regola d'arresto forte uniforme per i mescolamenti top-in-at-random impiega più di $k = \lceil n \log n + cn \rceil$ passi con probabilità bassa:

$$\text{Prob}\big[T > k\big] \;\leq\; e^{-c}.$$

Questo significa a sua volta che dopo $k = \lceil n \log n + cn \rceil$ mescolamenti top-in-at-random, il nostro mazzo è "vicino all'essere aleatorio", con

$$d(k) \;=\; \|\mathsf{Top}^{*k} - \mathsf{U}\| \;\leq\; e^{-c},$$

grazie al seguente lemma, semplice ma cruciale.

Lemma. *Sia $\mathsf{Q} : \mathfrak{S}_n \longrightarrow \mathbb{R}$ una distribuzione di probabilità che definisce un processo di mescolamento Q^{*k} con una regola d'arresto forte uniforme il cui tempo di arresto sia T. Allora per ogni $k \geq 0$,*

$$\|\mathsf{Q}^{*k} - \mathsf{U}\| \;\leq\; \text{Prob}\big[T > k\big].$$

■ **Dimostrazione.** Se X è una variabile aleatoria con valori in $\mathfrak{S}_n$, con distribuzione di probabilità Q, allora indichiamo con $\mathsf{Q}(S)$ la probabilità che X prenda un valore in $S \subseteq \mathfrak{S}_n$. Dunque $\mathsf{Q}(S) = \text{Prob}[X \in S]$ e nel caso della distribuzione uniforme $\mathsf{Q} = \mathsf{U}$ otteniamo

$$\mathsf{U}(S) \;=\; \text{Prob}\big[X \in S\big] \;=\; \frac{|S|}{n!}.$$

Per ogni sottoinsieme $S \subseteq \mathfrak{S}_n$, otteniamo la probabilità che dopo k passi il nostro mazzo sia ordinato in base ad una permutazione in S come

$$\begin{aligned}
\mathsf{Q}^{*k}(S) &= \text{Prob}[X_k \in S] \\
&= \sum_{j \le k} \text{Prob}[X_k \in S \ \wedge \ T = j] \ + \ \text{Prob}[X_k \in S \ \wedge \ T > k] \\
&= \sum_{j \le k} \mathsf{U}(S)\,\text{Prob}[T = j] \ + \ \text{Prob}[X_k \in S \,|\, T > k] \cdot \text{Prob}[T > k] \\
&= \mathsf{U}(S)\,(1 - \ \text{Prob}[T > k]) \ + \ \text{Prob}[X_k \in S \,|\, T > k] \cdot \text{Prob}[T > k] \\
&= \mathsf{U}(S) \ + \ \big(\text{Prob}[X_k \in S \,|\, T > k] - \mathsf{U}(S)\big) \cdot \text{Prob}[T > k].
\end{aligned}$$

Ciò dà

$$|\mathsf{Q}^{*k}(S) - \mathsf{U}(S)| \ \le \ \text{Prob}[T > k]$$

poiché

$$\text{Prob}[X_k \in S \,|\, T > k] \ - \ \mathsf{U}(S)$$

è la differenza di due probabilità, pertanto il suo valore assoluto massimo vale al massimo 1. □

A questo punto abbiamo completato la nostra analisi del mescolamento top-in-at-random: abbiamo dimostrato la seguente maggiorazione per il numero di mescolamenti necessari per essere "vicino all'essere aleatorio".

Teorema 1. *Siano $c \ge 0$ e $k := \lceil n \log n + cn \rceil$. Allora dopo aver effettuato k mescolamenti top-in-at-random su un mazzo di n carte, la distanza in variazione dalla distribuzione uniforme soddisfa*

$$d(k) \ := \ \|\mathsf{Top}^{*k} - \mathsf{U}\| \ \le \ e^{-c}.$$

Si può anche verificare che la distanza in variazione $d(k)$ rimane grande se effettuiamo assai meno di $n \log n$ mescolamenti top-in-at-random. La ragione è che un numero inferiore di mescolamenti non basterà a distruggere l'ordine relativo delle carte più in basso nel mazzo.

Ovviamente, i mescolamenti top-in-at-random sono estremamente inefficaci; con le maggiorazioni del nostro teorema, servono circa $n \log n \approx 205$ mescolamenti aleatori dall'alto perché un mazzo di $n = 52$ carte sia ben mescolato. Per tale ragione rivolgeremo ora la nostra attenzione ad un modello di mescolamento molto più interessante e realistico.

Mescolamenti all'americana

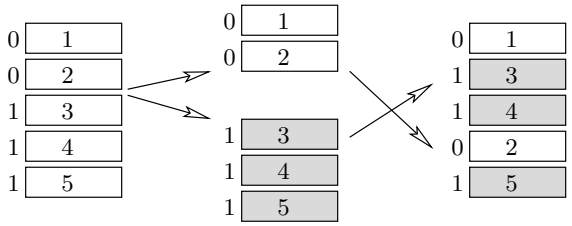
Questo è quello che fanno i croupier al casinò: prendono il mazzo, lo dividono in due parti, e queste sono quindi inserite l'una nell'altra, ad esempio facendo cadere le carte dal fondo delle due metà seguendo un andamento irregolare.

Un mescolamento all'americana

Di nuovo un riffle shuffle determina una certa permutazione delle carte nel mazzo, che supponiamo siano inizialmente contrassegnate da 1 a n, essendo la carta in cima la numero 1. I mescolamenti all'americana corrispondono esattamente alle permutazioni $\pi \in \mathfrak{S}_n$ tali che la successione

$$(\pi(1), \pi(2), \ \ldots \ , \pi(n))$$

consista in due successioni intersecate crescenti (solo per la permutazione identità si tratta di una sola successione crescente) e che vi siano esattamente $2^n - n$ mescolamenti all'americana distinti in un mazzo di n carte.



In effetti, se il gruppo è diviso in modo che le t carte in cima $(0 \leq t \leq n)$ siano prese nella mano destra e le altre $n-t$ carte nella mano sinistra, allora vi sono $\binom{n}{t}$ modi di alternare le due mani, i quali generano permutazioni diverse, eccetto che per ogni t vi è una possibilità di ottenere la permutazione identità.

Ora non è chiaro quale distribuzione di probabilità vada assegnata ai mescolamenti all'americana; non vi è una risposta unica poiché i croupier (sia professionisti che non) mescolano diversamente. Tuttavia, il modello seguente, sviluppato inizialmente da Edgar N. Gilbert e Claude Shannon nel 1955 (presso il dipartimento a quei tempi leggendario di "Matematica della Comunicazione" dei Bell Labs), ha diverse virtù:

- è elegante, semplice e sembra naturale;
- modella piuttosto bene il modo in cui un dilettante effettuerebbe i mescolamenti all'americana;
- abbiamo la possibilità di analizzarlo.

Ecco tre descrizioni diverse della stessa probabilità di distribuzione Rif su $\mathfrak{S}_n$:

1. $\mathsf{Rif} : \mathfrak{S}_n \longrightarrow \mathbb{R}$ è definito da

$$\mathsf{Rif}(\pi) \ := \ \begin{cases} \frac{n+1}{2^n} & \text{se } \pi = \mathrm{id}, \\ \frac{1}{2^n} & \text{se } \pi \text{ consiste in due successioni crescenti}, \\ 0 & \text{negli altri casi.} \end{cases}$$

2. Togliete t carte dal mazzo con probabilità $\frac{1}{2^n}\binom{n}{t}$, prendetele nella mano destra e prendete il resto del mazzo nella mano sinistra. Ora quando avrete r carte nella mano destra e ℓ nella sinistra, fate cadere la carta in fondo dalla vostra mano destra con probabilità $\frac{r}{r+\ell}$, e dalla sinistra con probabilità $\frac{\ell}{r+\ell}$. Ora ripetete!

3. Un *mescolamento inverso* prenderebbe un sottoinsieme delle carte nel mazzo, le rimuoverebbe e le metterebbe in cima alle carte restanti, mantenendo l'ordine relativo in entrambe le parti del mazzo. Tale movimento è determinato dal sottoinsieme delle carte: prendete tutti i sottoinsiemi con uguale probabilità.

   Equivalentemente, si contrassegni ogni carta con "0" o "1", arbitrariamente ed indipendentemente con probabilità $\frac{1}{2}$, e si spostino le carte contrassegnate con "0" in cima.

I mescolamenti inversi corrispondono alle permutazioni $\pi = (\pi(1), \ldots, \pi(n))$ che sono crescenti ad eccezione di al massimo una "discesa". (Solo la permutazione identità non ha discesa.)

È semplice vedere che queste descrizioni danno le stesse distribuzioni di probabilità. Per verificare che (1) $\Longleftrightarrow$ (3) si osservi semplicemente che otteniamo la permutazione identità ogniqualvolta tutte le carte contrassegnate con 0 sono sopra tutte quelle contrassegnate con 1.

Questo definisce il modello. Ma come possiamo analizzarlo? Quanti mescolamenti all'americana sono necessari per avvicinarsi all'ordine aleatorio? Non avremo esattamente la migliore risposta possibile, ma una piuttosto buona, combinando tre componenti:

(1) analizziamo piuttosto i mescolamenti all'americana inversi;

(2) descriviamo una regola d'arresto forte uniforme per questi;

(3) dimostriamo che la chiave per analizzarla è data dal paradosso del compleanno!

Teorema 2. *Dopo aver effettuato k mescolamenti all'americana su un mazzo di n carte, la distanza in variazione da una distribuzione uniforme soddisfa*

$$\|\mathsf{Rif}^{*k} - \mathsf{U}\| \;\leq\; 1 - \prod_{i=1}^{n-1}\left(1 - \frac{i}{2^k}\right).$$

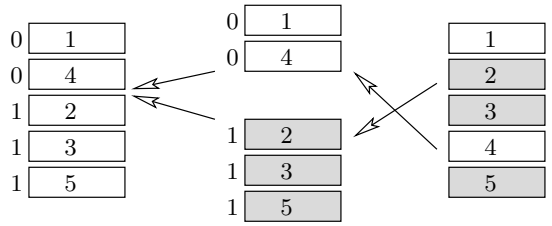
■ **Dimostrazione.** (1) Possiamo effettivamente analizzare i mescolamenti all'americana inversi e provare a vedere quanto rapidamente ci portano dalla distribuzione iniziale a quella (quasi) uniforme. Questi mescolamenti inversi corrispondono alla distribuzione di probabilità data da $\overline{\mathsf{Rif}}(\pi) := \mathsf{Rif}(\pi^{-1})$.

Ora il fatto che ogni permutazione abbia un'unica inversa ed il fatto che $\mathsf{U}(\pi) = \mathsf{U}(\pi^{-1})$, danno

$$\|\mathsf{Rif}^{*k} - \mathsf{U}\| \;=\; \|\,\overline{\mathsf{Rif}}^{*k} - \mathsf{U}\|.$$

(Questo è il lemma di inversione di Reeds!)

(2) In ogni riffle shuffle inverso, ogni carta è associata ad una cifra 0 o 1:

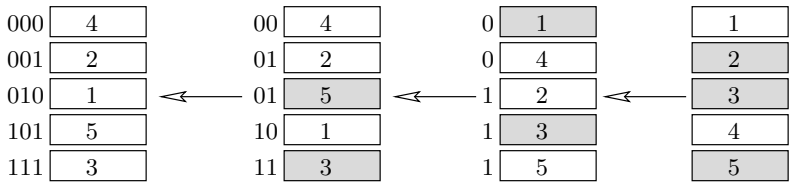


Se ricordiamo queste cifre — supponiamo di scriverle sulle carte — allora dopo k mescolamenti all'americana inversi, ogni carta ha avuto una striscia ordinata di k cifre. La nostra regola d'arresto è:

"STOP non appena tutte le carte abbiano strisce diverse."

Quando questo si verifica, le carte nel mazzo sono *ordinate* in base ai numeri binari $b_k b_{k-1} \dots b_2 b_1$, dove b_i è il bit [NdT: la cifra binaria] che la carta ha preso nell'i-esimo riffle shuffle inverso. Poiché questi bit sono perfettamente arbitrari e indipendenti, questa regola d'arresto è forte uniforme!

Nell'esempio che segue, per $n = 5$ carte, ci servono $T = 3$ mescolamenti inversi prima dell'arresto:



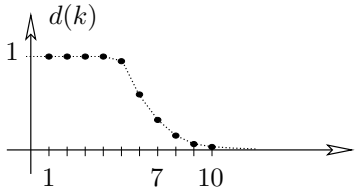
(3) Il tempo atteso T impiegato da questa regola d'arresto è distribuito in base al paradosso del compleanno, con $K = 2^k$: mettiamo due carte nella stessa casella se hanno lo stesso contrassegno $b_k b_{k-1} \dots b_2 b_1 \in \{0,1\}^k$. Allora ci sono $K = 2^k$ caselle e la probabilità che una casella abbia più di una carta è

$$\text{Prob}[T > k] \;=\; 1 - \prod_{i=1}^{n-1} \left(1 - \frac{i}{2^k}\right),$$

e come abbiamo visto ciò limita la distanza della variazione $\|\mathsf{Rif}^{*k} - \mathsf{U}\| = \|\,\overline{\mathsf{Rif}}^{*k} - \mathsf{U}\|$. □

k	$d(k)$
1	1.000
2	1.000
3	1.000
4	1.000
5	0.952
6	0.614
7	0.334
8	0.167
9	0.085
10	0.043

La distanza in variazione dopo k mescolamenti all'americana, secondo [2]



Quante volte dobbiamo mescolare dunque? Per n grandi servono circa $k = 2\log_2(n)$ mescolamenti. In effetti, ponendo $k := 2\log_2(cn)$ per un dato $c \geq 1$ troviamo (con un po' di analisi elementare) che $P[T > k] \approx 1 - e^{-\frac{1}{2c^2}} \approx \frac{1}{2c^2}$. In modo esplicito, per $n = 52$ carte la maggiorazione del Teorema 2 fornisce $d(10) \leq 0.73$, $d(12) \leq 0.28$, $d(14) \leq 0.08$, quindi $k = 12$ dovrebbe essere "abbastanza aleatorio" nella pratica. Ma non facciamo 12 mescolamenti "in pratica"— e non sono veramente necessari, come mostra un'analisi più dettagliata (con i risultati riportati a margine). L'analisi dei mescolamenti all'americana è parte di un'accesa discussione

in corso sulla giusta misura di cosa sia "abbastanza aleatorio". Diaconis [4] fornisce una guida agli sviluppi recenti.

Tutto questo è importante? Sì, lo è: anche dopo tre buoni mescolamenti all'americana un mazzo ordinato di 52 carte sembra piuttosto aleatorio . . . ma non lo è. Martin Gardner [5, Capitolo 7] descrive un certo numero di trucchi notevoli con le carte basati sull'ordine nascosto in un mazzo del genere!

Bibliografia

[1] D. ALDOUS & P. DIACONIS: *Shuffling cards and stopping times,* Amer. Math. Monthly **93** (1986), 333-348.

[2] D. BAYER & P. DIACONIS: *Trailing the dovetail shuffle to its lair,* Annals Applied Probability **2** (1992), 294-313.

[3] E. BEHRENDS: *Introduction to Markov Chains,* Vieweg, Braunschweig/ Wiesbaden 2000.

[4] P. DIACONIS: *Mathematical developments from the analysis of riffle shuffling,* in: "Groups, Combinatorics and Geometry. Durham 2001" (A. A. Ivanov, M. W. Liebeck and J. Saxl, eds.), World Scientific, Singapore 2003, pp. 73-97.

[5] M. GARDNER: *Mathematical Magic Show,* Knopf, New York/Allen & Unwin, London 1977.

[6] E. N. GILBERT: *Theory of Shuffling,* Technical Memorandum, Bell Laboratories, Murray Hill NJ, 1955.

Sufficientemente a caso?

Cammini su reticoli e determinanti — Capitolo 25

L'essenza stessa della matematica è dimostrare teoremi; in effetti, è questo ciò che fanno i matematici: dimostrano teoremi. Ma a dire il vero, quello che vale la pena di dimostrare, una volta nella vita, è un *Lemma*, come il Lemma di Fatou nell'analisi, il Lemma di Gauss nella teoria dei numeri, o il Lemma di Burnside-Frobenius nel calcolo combinatorio.

Ora, che cosa rende un'affermazione matematica un Lemma celebre? In primo luogo, essa dovrebbe essere applicabile ad un'ampia gamma di casi, così come a problemi apparentemente non correlati. In secondo luogo, una volta compreso, il risultato dovrebbe essere completamente ovvio. La reazione del lettore potrebbe essere un puro scatto d'invidia: "Perché non me ne sono accorto prima?". In terzo luogo, da un punto di vista estetico, il Lemma — ed in particolare la sua dimostrazione — dovrebbe essere splendido!

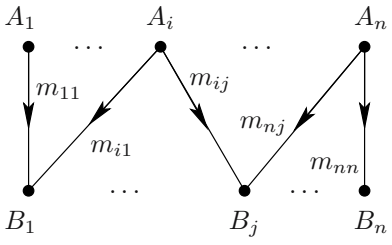
In questo capitolo consideriamo una di tali meraviglie di ragionamento matematico, un lemma di numerazione che apparve per la prima volta in una pubblicazione di Bernt Lindström nel 1972. Largamente sottovalutato all'epoca, il risultato divenne all'improvviso un classico nel 1985, quando Ira Gessel e Gerard Viennot lo riscoprirono, e dimostrarono in una meravigliosa pubblicazione come il lemma potesse essere applicato con successo ad una varietà di difficili problemi di enumerazione.

Il punto di partenza è la consueta rappresentazione del determinante di una matrice sotto forma di permutazione. Sia $M = (m_{ij})$ una matrice reale di dimensione $n \times n$. Allora

$$\det M \;=\; \sum_{\sigma} \operatorname{sign}\sigma \; m_{1\sigma(1)}\, m_{2\sigma(2)} \cdots m_{n\sigma(n)}, \qquad (1)$$

dove σ si estende su tutte le permutazioni di $\{1, 2, \ldots, n\}$ ed il segno di σ è 1 o -1, a seconda che σ sia il prodotto di un numero pari o dispari di trasposizioni.

Ora passiamo ai grafi, più precisamente ai grafi *bipartiti orientati e pesati*. Si usino i vertici $A_1, \ldots, A_n$ per rappresentare le righe di M, e $B_1, \ldots, B_n$ per rappresentare le colonne. Per ogni coppia (i, j) si tracci un arco da A_i a B_j e gli si attribuisca il peso m_{ij}, come indicato in figura.



Se riferita a questo grafo, la formula (1) ha la seguente interpretazione:

- La parte sinistra è il determinante della *matrice di adiacenza* M, il cui elemento (i, j) è il *peso* dell' unico cammino orientato da A_i a B_j.

- La parte destra è la somma pesata (segnata) su tutti i *sistemi dei cammini a vertici disgiunti* da $\mathcal{A} = \{A_1, \ldots, A_n\}$ a $\mathcal{B} = \{B_1, \ldots, B_n\}$. Tale sistema $\mathcal{P}_\sigma$ è determinato dai cammini

$$A_1 \to B_{\sigma(1)}, \ \ldots \ , A_n \to B_{\sigma(n)},$$

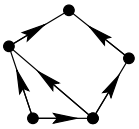
ed il *peso* del sistema dei cammini $\mathcal{P}_\sigma$ è il prodotto dei pesi di ogni cammino:

$$w(\mathcal{P}_\sigma) \ = \ w(A_1 \to B_{\sigma(1)}) \ \cdots \ w(A_n \to B_{\sigma(n)}).$$

Con questa interpretazione la formula (1) si legge

$$\det M \ = \ \sum_\sigma \operatorname{sign}\sigma \, w(\mathcal{P}_\sigma).$$

E qual è il risultato di Gessel e Viennot? È la generalizzazione naturale di (1) da grafi bipartiti a grafi arbitrari. È precisamente questa caratteristica che rende il Lemma così ampiamente applicabile — e per di più, la sua dimostrazione è prodigiosamente semplice ed elegante.



Un grafo direzionale aciclico

Cominciamo a mettere insieme i concetti necessari. Abbiamo un grafo direzionale aciclico $G = (V, E)$, dove *aciclico* significa che non vi sono cicli direzionali in G. In particolare, vi è solo un numero finito di cammini orientati tra ogni coppia di vertici A e B; includiamo tutti i cammini banali $A \to A$ di lunghezza 0. Ogni arco e ha un peso $w(e)$. Se P è un cammino orientato da A a B, che indicheremo per brevità con $P : A \to B$, allora definiamo il *peso* di P come

$$w(P) \ := \ \prod_{e \in P} w(e).$$

Grazie alla definizione $w(P) = 1$ se P è un cammino di lunghezza 0.

Ora siano $\mathcal{A} = \{A_1, \ldots, A_n\}$ e $\mathcal{B} = \{B_1, \ldots, B_n\}$ due insiemi di n vertici, dove $\mathcal{A}$ e $\mathcal{B}$ non devono essere necessariamente disgiunti. Ad $\mathcal{A}$ e $\mathcal{B}$ associamo la *matrice del cammino* $M = (m_{ij})$ tale che

$$m_{ij} \ := \ \sum_{P : A_i \to B_j} w(P).$$

Un *sistema dei cammini* $\mathcal{P}$ *da* $\mathcal{A}$ *a* $\mathcal{B}$ consiste in una permutazione σ e in n cammini $P_i : A_i \to B_{\sigma(i)}$, per $i = 1, \ldots, n$; scriviamo $\operatorname{sign} \mathcal{P} = \operatorname{sign}\sigma$. Il *peso* di $\mathcal{P}$ è il prodotto dei pesi dei cammini

$$w(\mathcal{P}) \ = \ \prod_{i=1}^{n} w(P_i), \tag{2}$$

ovvero il prodotto dei pesi di tutti gli archi del sistema dei cammini.

Infine, diciamo che il sistema dei cammini $\mathcal{P} = (P_1, \ldots, P_n)$ è *a vertici disgiunti* se i cammini di $\mathcal{P}$ sono a vertici disgiunti a due a due.

Lemma. *Siano $G = (V, E)$ un grafo aciclico finito, orientato e pesato $\mathcal{A} = \{A_1, \ldots, A_n\}$ e $\mathcal{B} = \{B_1, \ldots, B_n\}$ due insiemi di n vertici e M la matrice dei cammini da $\mathcal{A}$ a $\mathcal{B}$. Allora*

$$\det M \;=\; \sum_{\substack{\mathcal{P} \text{ sistema dei cammini} \\ \text{a vertici disgiunti}}} \operatorname{sign} \mathcal{P}\; w(\mathcal{P}). \tag{3}$$

■ **Dimostrazione.** Un tipico addendo di $\det(M)$ è $\operatorname{sign}\sigma\; m_{1\sigma(1)} \cdots m_{n\sigma(n)}$, che si può scrivere come

$$\operatorname{sign}\sigma\;\Big(\sum_{P_1 : A_1 \to B_{\sigma(1)}} w(P_1)\Big)\;\cdots\;\Big(\sum_{P_n : A_n \to B_{\sigma(n)}} w(P_n)\Big).$$

Sommando su σ troviamo immediatamente da (2) che

$$\det M \;=\; \sum_{\mathcal{P}} \operatorname{sign}\mathcal{P}\; w(\mathcal{P}),$$

dove $\mathcal{P}$ si estende su *tutti* i sistemi dei cammini da $\mathcal{A}$ a $\mathcal{B}$ (che siano o meno a vertici disgiunti). Pertanto per arrivare a (3) dobbiamo semplicemente dimostrare che

$$\sum_{\mathcal{P} \in N} \operatorname{sign}\mathcal{P}\; w(\mathcal{P}) \;=\; 0\,, \tag{4}$$

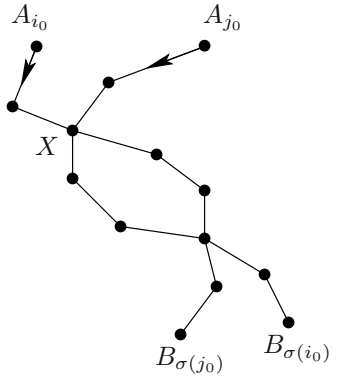
dove N è l'insieme di tutti i sistemi dei cammini che *non* siano a vertici disgiunti. Questo si ottiene mediante un'argomentazione di singolare bellezza. In effetti, esibiamo un'involuzione $\pi : N \to N$ (senza punti fissi) tale che per $\mathcal{P}$ e $\pi\mathcal{P}$

$$w(\pi\mathcal{P}) \;=\; w(\mathcal{P}) \qquad \text{e} \qquad \operatorname{sign}\pi\mathcal{P} \;=\; -\operatorname{sign}\mathcal{P}\,.$$

Chiaramente, ciò implicherà (4) e dunque la formula (3) del Lemma.

L'involuzione π è definita in modo assolutamente naturale. Sia $\mathcal{P} \in N$ con cammini $P_i : A_i \to B_{\sigma(i)}$. Per definizione, alcune coppie di cammini si intersecheranno:

- Sia i_0 l'indice minimo tale che P_{i_0} condivida un vertice con un altro cammino.
- Sia X il primo vertice comune di questo tipo sul cammino P_{i_0}.
- Sia j_0 l'indice minimo ($j_0 > i_0$) tale che P_{j_0} abbia il vertice X in comune con P_{i_0}.



Costruiamo ora il nuovo sistema $\pi\mathcal{P} = (P'_1, \ldots, P'_n)$ come segue:

- Si ponga $P'_k = P_k$ per ogni $k \neq i_0, j_0$.
- Il nuovo cammino P'_{i_0} va da A_{i_0} a X lungo P_{i_0} e quindi continua verso $B_{\sigma(j_0)}$ lungo P_{j_0}. Analogamente, P'_{j_0} va da A_{j_0} a X lungo P_{j_0} e continua verso $B_{\sigma(i_0)}$ lungo P_{i_0}.

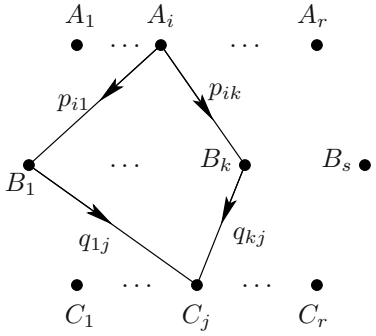
Chiaramente $\pi(\pi\mathcal{P}) = \mathcal{P}$, poiché l'indice i_0, il vertice X, e l'indice j_0 sono gli stessi di prima. In altre parole, applicando π due volte torniamo ai vecchi cammini P_i. In seguito, poiché $\pi\mathcal{P}$ e $\mathcal{P}$ usano esattamente gli stessi archi, abbiamo certamente $w(\pi\mathcal{P}) = w(\mathcal{P})$. Ed infine, poiché la nuova permutazione σ' si ottiene moltiplicando σ per la trasposizione (i_0, j_0), troviamo che $\operatorname{sign}\pi\mathcal{P} = -\operatorname{sign}\mathcal{P}$ e questo è tutto. □

Il Lemma di Gessel-Viennot può essere usato per derivare tutte le proprietà fondamentali dei determinanti, semplicemente analizzando i grafi appropriati. Consideriamo un esempio particolarmente notevole, la formula di Binet-Cauchy, che fornisce una generalizzazione molto utile della regola del prodotto per i determinanti.

Teorema. *Se P è una matrice $(r \times s)$ e Q una matrice $(s \times r)$, $r \le s$, allora*

$$\det(PQ) \;=\; \sum_{\mathcal{Z}} (\det P_{\mathcal{Z}})(\det Q_{\mathcal{Z}}),$$

dove $P_{\mathcal{Z}}$ è la sottomatrice $(r \times r)$ di P le cui colonne sono definite da $\mathcal{Z}$, mentre $Q_{\mathcal{Z}}$ è la sottomatrice $(r \times r)$ di Q costituita dalle corrispondenti righe $\mathcal{Z}$.



■ **Dimostrazione.** Come in precedenza, si faccia corrispondere a P il grafo bipartito su $\mathcal{A}$ e $\mathcal{B}$ ed analogamente facciamo corrispondere a Q il grafo bipartito su $\mathcal{B}$ e $\mathcal{C}$. Si consideri ora il grafo concatenato come indicato nella figura a sinistra, e si osservi che m_{ij}, l'elemento (i, j) della matrice del cammino M da $\mathcal{A}$ a $\mathcal{C}$ è esattamente $m_{ij} = \sum_k p_{ik}q_{kj}$, pertanto $M = PQ$. Poiché i sistemi dei cammini a vertici disgiunti da $\mathcal{A}$ a $\mathcal{C}$ nel grafo concatenato corrispondono a coppie di sistemi da $\mathcal{A}$ a $\mathcal{Z}$ e rispettivamente da $\mathcal{Z}$ a $\mathcal{C}$, il risultato si deduce immediatamente dal Lemma, notando che $\operatorname{sign}(\sigma\tau) = (\operatorname{sign}\sigma)\cdot(\operatorname{sign}\tau)$. □

Il Lemma di Gessel-Viennot è anche la fonte di un gran numero di risultati che mettono in relazione i determinanti con le proprietà enumerative. La ricetta è sempre la stessa: si interpreti la matrice M come una matrice di cammini e si provi a calcolare il membro di destra di (3). Come illustrazione considereremo il problema originale studiato da Gessel e Viennot, che li condusse al loro Lemma:

> *Siano $a_1 < a_2 < \ldots < a_n$ e $b_1 < b_2 < \ldots < b_n$ due insiemi di numeri naturali. Vogliamo calcolare il determinante della matrice $M = (m_{ij})$, essendo m_{ij} il coefficiente binomiale $\binom{a_i}{b_j}$.*

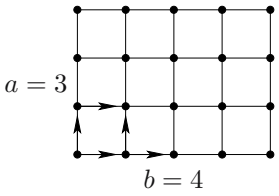
1							
1	1						
1	2	1					
1	**3**	3	**1**				
1	**4**	6	**4**	**1**			
1	5	10	10	5	1		
1	**6**	15	**20**	**15**	6	1	
1	7	21	35	35	21	7	1

In altre parole, Gessel e Viennot cercavano i determinanti di matrici quadrate arbitrarie estratte dal triangolo di Pascal, come ad esempio la matrice

$$\det\begin{pmatrix} \binom{3}{1} & \binom{3}{3} & \binom{3}{4} \\ \binom{4}{1} & \binom{4}{3} & \binom{4}{4} \\ \binom{6}{1} & \binom{6}{3} & \binom{6}{4} \end{pmatrix} \;=\; \det\begin{pmatrix} 3 & 1 & 0 \\ 4 & 4 & 1 \\ 6 & 20 & 15 \end{pmatrix}$$

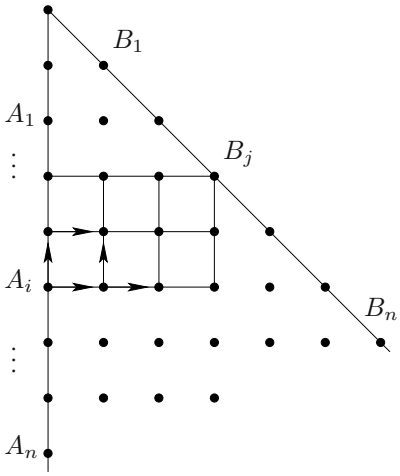
data dagli elementi in grassetto del triangolo di Pascal mostrati a margine.

Come passo preliminare verso la soluzione del problema ricordiamo un ben noto risultato che collega i coefficienti binomiali ai cammini sui reticoli. Consideriamo un reticolo $a \times b$ come quello rappresentato a margine. Allora se i soli passi consentiti per i cammini sono verso l'alto (Nord) e verso destra (Est) il numero di cammini dall'angolo in basso a sinistra a quello in alto a destra è $\binom{a+b}{a}$.



La dimostrazione è semplice: ogni cammino consiste in una successione arbitraria di b passi verso "est" e a passi verso "nord" e pertanto può essere codificato con una successione della forma NENEEEN, che consiste in $a+b$ lettere, di cui a N e b E. Il numero di tali catene è il numero dei modi di scegliere a posizioni di lettere N su un totale di $a + b$ posizioni, che è $\binom{a+b}{a} = \binom{a+b}{b}$.

Si guardi ora la figura a margine, dove A_i è il punto $(0, -a_i)$ e B_j il punto $(b_j, -b_j)$.



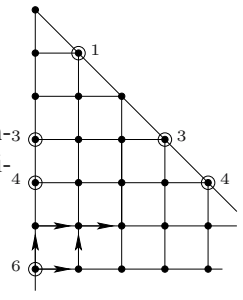
Il numero di cammini da A_i a B_j su questa griglia (i quali utilizzano solo passi verso nord e verso est) è, sulla base di quanto abbiamo appena dimostrato, $\binom{b_j+(a_i-b_j)}{b_j} = \binom{a_i}{b_j}$. In altre parole, la matrice M dei coefficienti binomiali è esattamente la matrice dei cammini da $\mathcal{A}$ a $\mathcal{B}$ nel grafo del reticolo orientato nel quale tutti gli archi hanno peso 1 e tutti gli archi sono diretti a nord o a est. Pertanto per calcolare $\det M$ possiamo applicare il Lemma di Gessel-Viennot. Una rapida riflessione mostra che ogni sistema dei cammini a vertici disgiunti $\mathcal{P}$ da $\mathcal{A}$ a $\mathcal{B}$ deve consistere in cammini $P_i : A_i \to B_i$ per ogni i. Pertanto la sola permutazione possibile è l'identità, che ha segno $= 1$, ed otteniamo lo splendido risultato

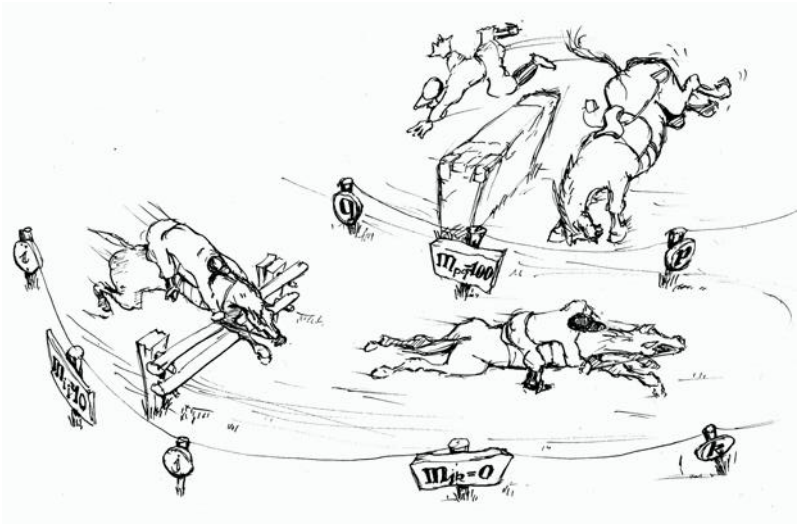
$$\det\left(\binom{a_i}{b_j}\right) \;=\; \#\text{ sistema dei cammini a vertici disgiunti da } \mathcal{A} \text{ a } \mathcal{B}.$$

In particolare, ciò implica il fatto tutt'altro che ovvio che $\det M$ è sempre non negativo, poiché il membro a destra dell'uguaglianza *conta* qualche cosa. Più precisamente, si ottiene dal Lemma Gessel-Viennot che $\det M = 0$ se e solo se $a_i < b_i$ per un dato i.

Nel piccolo esempio precedente,

$$\det\begin{pmatrix} \binom{3}{1} & \binom{3}{3} & \binom{3}{4} \\ \binom{4}{1} & \binom{4}{3} & \binom{4}{4} \\ \binom{6}{1} & \binom{6}{3} & \binom{6}{4} \end{pmatrix} \;=\; \#\begin{array}{l}\text{sistema dei cam-}\\ \text{mini a vertici di-}\\ \text{sgiunti}\end{array}$$





"Cammini sui reticoli"

Bibliografia

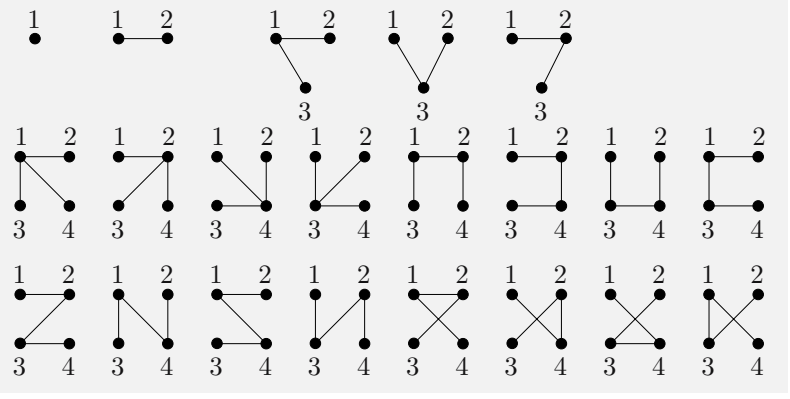
[1] I. M. GESSEL & G. VIENNOT: *Binomial determinants, paths, and hook length formulae,* Advances in Math. **58** (1985), 300-321.

[2] B. LINDSTRÖM: *On the vector representation of induced matroids,* Bulletin London Math. Soc. **5** (1973), 85-90.

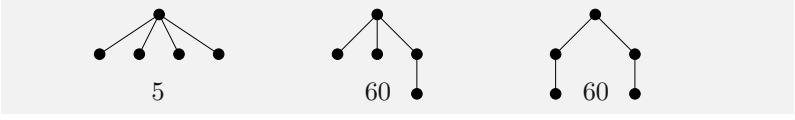
La formula di Cayley per il numero di alberi

Capitolo 26

Una delle formule più belle nel calcolo combinatorio riguarda il numero di alberi etichettati. Si consideri l'insieme $N = \{1, 2, \ldots, n\}$. Quanti alberi diversi possiamo formare su questo insieme di vertici? Indichiamo tale numero con T_n. La conta "manuale" dà $T_1 = 1$, $T_2 = 1$, $T_3 = 3$, $T_4 = 16$, con gli alberi mostrati nella tabella seguente:



Si noti che consideriamo alberi *etichettati*, pertanto, sebbene vi sia solo un albero di ordine 3 nel senso dell'isomorfismo fra grafi, vi sono 3 alberi etichettati diversi ottenuti contrassegnando il vertice più interno con 1, 2 o 3. Per $n = 5$ vi sono tre alberi non isomorfi:

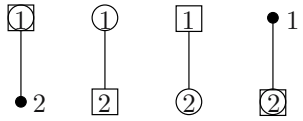


Per il primo albero vi sono chiaramente 5 modi differenti di etichettare e per il secondo ed il terzo vi sono $\frac{5!}{2} = 60$ modi, pertanto otteniamo $T_5 = 125$. Questo dovrebbe essere sufficiente per congetturare che $T_n = n^{n-2}$ e tale è esattamente il risultato di Cayley.

Teorema. *Vi sono n^{n-2} alberi etichettati distinti su n vertici.*

Questa splendida formula si presta a dimostrazioni altrettanto splendide, basate su una varietà di tecniche algebriche e combinatorie. Ne introdurremo tre prima di presentare la dimostrazione che a tutt'oggi è la migliore di tutte.

Arthur Cayley



I quattro alberi di $\mathcal{T}_2$

■ **Prima dimostrazione (Biiezione).** Il metodo classico e più diretto consiste nel trovare una biiezione fra l'insieme di tutti gli alberi su n vertici e un altro insieme la cui cardinalità sia notoriamente n^{n-2}. Naturalmente, viene in mente l'insieme di tutte le successioni ordinate $(a_1, \ldots, a_{n-2})$ tali che $1 \leq a_i \leq n$. Pertanto vogliamo codificare univocamente ogni albero T con una successione $(a_1 \ldots, a_{n-2})$. Tale codice fu trovato da Prüfer ed è presente in quasi tutti i libri di teoria dei grafi.

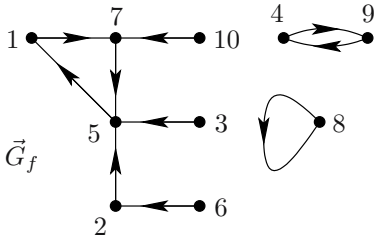
Qui vogliamo discutere un'altra dimostrazione di biiezione, dovuta a Joyal, meno nota ma di uguale eleganza e semplicità. Per questo, consideriamo non solo degli alberi t su $N = \{1, \ldots, n\}$ ma alberi insieme con due vertici distinti, l'*estremo sinistro* ◯ e l'*estremo destro* □, che possono coincidere. Sia $\mathcal{T}_n = \{(t; ◯, □)\}$ questo nuovo insieme; allora, chiaramente, $|\mathcal{T}_n| = n^2 T_n$.

Il nostro obiettivo è pertanto dimostrare che $|\mathcal{T}_n| = n^n$. Ora vi è un insieme la cui dimensione è notoriamente n^n, precisamente l'insieme N^N di tutte le trasformazioni di N in N. Pertanto la nostra formula è dimostrata se troviamo una biiezione da N^N su $\mathcal{T}_n$.

Sia $f : N \longrightarrow N$ una trasformazione qualunque. Rappresentiamo f come un grafo orientato $\vec{G}_f$ tracciando archi da i a $f(i)$.

Ad esempio, la trasformazione

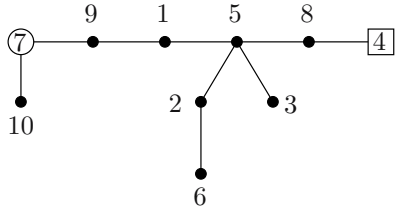
$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 7 & 5 & 5 & 9 & 1 & 2 & 5 & 8 & 4 & 7 \end{pmatrix}$$



è rappresentata dal grafo orientato a margine.

Si consideri un componente di $\vec{G}_f$. Poiché vi è esattamente un arco a partire da ogni vertice, il componente contiene un numero uguale di vertici e di archi e pertanto esattamente un ciclo orientato. Sia $M \subseteq N$ l'unione degli insiemi dei vertici di tali cicli. Una breve riflessione mostra che M è l'*unico* sottoinsieme massimale di N tale che la restrizione di f su M si comporti come una biiezione su M. Si scriva $f|_M = \begin{pmatrix} a & b & \ldots & z \\ f(a) & f(b) & \ldots & f(z) \end{pmatrix}$ in modo che i numeri $a, b, \ldots, z$ nella prima riga appaiano nell'ordine naturale. In questo modo abbiamo un ordinamento $f(a), f(b), \ldots, f(z)$ di M in base alla seconda riga. Ora $f(a)$ è il nostro estremo sinistro e $f(z)$ è quello destro.

L'albero t corrispondente alla trasformazione f si costruisce ora come segue: si tracci $f(a), \ldots, f(z)$ in quest'ordine come un *cammino* da $f(a)$ a $f(z)$ e si riempiano i vertici rimanenti come fatto in $\vec{G}_f$ (eliminando le frecce).



Nel nostro esempio precedente otteniamo $M = \{1, 4, 5, 7, 8, 9\}$

$$f|_M = \begin{pmatrix} 1 & 4 & 5 & 7 & 8 & 9 \\ 7 & 9 & 1 & 5 & 8 & 4 \end{pmatrix}$$

e pertanto l'albero t mostrato a margine.

È immediato capire come invertire questa corrispondenza: dato un albero t, consideriamo l'unico cammino P dall'estremo sinistro a quello destro. Otteniamo così l'insieme M e la trasformazione $f|_M$. Le corrispondenze rimanenti $i \to f(i)$ sono quindi "riempite" in base ai cammini unici da i a P. □

■ **Seconda dimostrazione (Algebra lineare).** Possiamo pensare a T_n come al numero di alberi generatori nel grafo completo K_n. Ora esaminiamo un grafo semplice connesso arbitrario G su $V = \{1, 2, \ldots, n\}$, indicando con $t(G)$ il numero di alberi generatori; pertanto $T_n = t(K_n)$. Il seguente celebre risultato è il *teorema della matrice di un albero* di Kirchhoff (si veda [1]). Si consideri la matrice di incidenza $B = (b_{ie})$ di G, le cui righe siano contrassegnate da V, le colonne da E, avendo definito $b_{ie} = 1$ o 0 a seconda che $i \in e$ oppure $i \notin e$. Si noti che $|E| \geq n - 1$ poiché G è connesso. In ogni colonna sostituiamo uno dei due 1 con -1 in modo arbitrario (ciò si traduce in un orientamento di G) e chiamiamo la nuova matrice C. $M = CC^T$ è allora una matrice simmetrica $(n \times n)$ con i gradi $d_1, \ldots, d_n$ sulla diagonale principale.

Proposizione. *Abbiamo $t(G) = \det M_{ii}$ per ogni $i = 1, \ldots, n$, dove M_{ii} risulta da M eliminando la i-esima riga e la i-esima colonna.*

■ **Dimostrazione.** La chiave per la dimostrazione è il teorema di Binet-Cauchy dimostrato nel capitolo precedente: quando P è una matrice $(r \times s)$ e Q una matrice $(s \times r)$, $r \leq s$, allora $\det(PQ)$ è uguale alla somma dei prodotti dei determinanti delle sottomatrici $(r \times r)$ corrispondenti, dove "corrispondenti" significa che prendiamo gli stessi indici per le r colonne di P e le r righe di Q.

Per M_{ii} ciò significa che

$$\det M_{ii} \;=\; \sum\nolimits_N \det N \cdot \det N^T \;=\; \sum\nolimits_N (\det N)^2,$$

dove N percorre tutte le sottomatrici $(n-1) \times (n-1)$ di $C \backslash \{\text{riga } i\}$. Le $n - 1$ colonne di N corrispondono ad un sottografo di G con $n - 1$ archi su n vertici, e resta da dimostrare che

$$\det N = \begin{cases} \pm 1 & \text{se questi archi generano un albero} \\ 0 & \text{altrimenti.} \end{cases}$$

Supponiamo che $n - 1$ archi non generino un albero. Allora vi è una componente che non contiene i. Poiché le righe corrispondenti di questa componente hanno somma 0, induciamo che esse sono linearmente dipendenti, e pertanto $\det N = 0$.

Si supponga ora che le colonne di N generino un albero. Allora vi è un vertice $j_1 \neq i$ di grado 1; sia e_1 l'arco incidente. Eliminando j_1, e_1 otteniamo un albero con $n - 2$ archi. Di nuovo vi è un vertice $j_2 \neq i$ di grado 1 con arco incidente e_2. Si continui in questo modo finché siano determinati $j_1, j_2, \ldots, j_{n-1}$ e $e_1, e_2, \ldots, e_{n-1}$ con $j_i \in e_i$. Ora si permutino le righe e le colonne per portare j_k nella riga k-esima ed e_k nella colonna k-esima.

"Un metodo non standard di contare gli alberi: mettete un gatto su ogni albero, fate fare una passeggiata al vostro cane, e contate quante volte abbaia"

Poiché per costruzione $j_k \notin e_\ell$ per $k < \ell$, vediamo che la nuova matrice N' è triangolare inferiore con tutti gli elementi sulla diagonale principale uguali a ± 1. Pertanto $\det N = \pm \det N' = \pm 1$ ed abbiamo terminato.

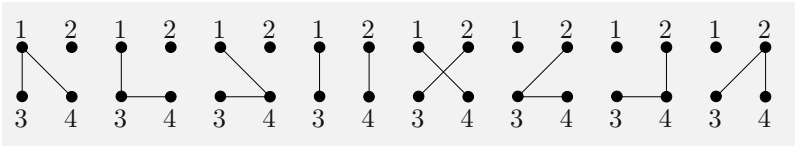
Per il caso speciale $G = K_n$ otteniamo chiaramente

$$M_{ii} = \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & & -1 \\ \vdots & & \ddots & \vdots \\ -1 & -1 & \dots & n-1 \end{pmatrix}$$

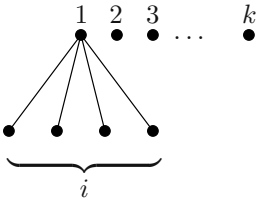
ed un semplice calcolo mostra che $\det M_{ii} = n^{n-2}$. □

■ **Terza dimostrazione (per induzione).** Un altro metodo classico nel calcolo combinatorio enumerativo consiste nello stabilire una relazione ricorsiva e di risolverla per induzione. L'idea riportata di seguito è dovuta essenzialmente a Riordan e Rényi. Per trovare la relazione appropriata, consideriamo un problema più generale (che appare già nella pubblicazione di Cayley). Sia A un k-insieme arbitrario dei vertici. Indichiamo con $T_{n,k}$ il numero di foreste (etichettate) su $\{1, \dots, n\}$ che consistono in k alberi dove i vertici di A appaiono in alberi diversi. Chiaramente, non conta l'insieme A, ma solo la dimensione k. Si noti che $T_{n,1} = T_n$.

Ad esempio, $T_{4,2} = 8$ per $A = \{1, 2\}$



Si consideri la foresta F con $A = \{1, 2, \dots, k\}$ e si supponga che 1 sia adiacente a i vertici, come indicato a margine. Eliminando 1, gli i vicini insieme con $2, \dots, k$ danno un vertice ciascuno nei componenti di una foresta che consiste in $k - 1 + i$ alberi. Poiché possiamo (ri)costruire F fissando dapprima i, quindi scegliendo gli i vicini di 1 ed infine la foresta $F \backslash 1$, otteniamo



$$T_{n,k} \;=\; \sum_{i=0}^{n-k} \binom{n-k}{i} T_{n-1,k-1+i} \tag{1}$$

per ogni $n \geq k \geq 1$, dove poniamo $T_{0,0} = 1$, $T_{n,0} = 0$ per $n > 0$. Si noti che $T_{0,0} = 1$ è necessario per garantire $T_{n,n} = 1$.

Proposizione. *Abbiamo*

$$T_{n,k} \;=\; k\, n^{n-k-1} \tag{2}$$

e pertanto, in particolare,

$$T_{n,1} \;=\; T_n \;=\; n^{n-2}.$$

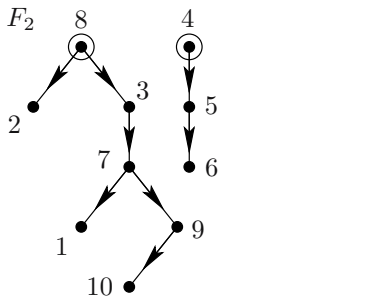
■ **Dimostrazione.** In base a (1), ed usando l'induzione, troviamo

$$\begin{aligned}
T_{n,k} &= \sum_{i=0}^{n-k} \binom{n-k}{i} (k-1+i)(n-1)^{n-1-k-i} \qquad (i \to n-k-i) \\
&= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1-i)(n-1)^{i-1} \\
&= \sum_{i=0}^{n-k} \binom{n-k}{i} (n-1)^{i} - \sum_{i=1}^{n-k} \binom{n-k}{i} i(n-1)^{i-1} \\
&= n^{n-k} - (n-k) \sum_{i=1}^{n-k} \binom{n-1-k}{i-1} (n-1)^{i-1} \\
&= n^{n-k} - (n-k) \sum_{i=0}^{n-1-k} \binom{n-1-k}{i} (n-1)^{i} \\
&= n^{n-k} - (n-k) n^{n-1-k} \;=\; k\, n^{n-1-k}. \qquad \square
\end{aligned}$$

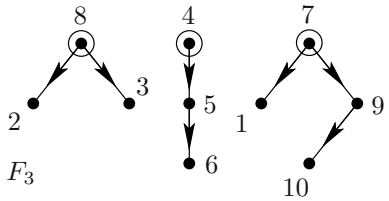
■ **Quarta dimostrazione (Conta doppia).** La meravigliosa idea che affrontiamo ora dovuta a Jim Pitman dà la formula di Cayley e la sua generalizzazione (2) senza induzione né biiezione; è semplicemente una conta intelligente in due sensi.

Una *foresta radicata* su $\{1, \ldots, n\}$ è una foresta con una radice scelta per ogni albero componente. Sia $\mathcal{F}_{n,k}$ l'insieme di tutte le foreste radicate che consistono in k alberi radicati. Pertanto $\mathcal{F}_{n,1}$ è l'insieme di tutti gli alberi radicati.

Si noti che $|\mathcal{F}_{n,1}| = nT_n$, poiché in ogni albero vi sono n scelte possibili per la radice. Consideriamo ora $F_{n,k} \in \mathcal{F}_{n,k}$ come un grafo *orientato* con tutti gli archi orientati a partire dalle radici. Diciamo che una foresta F *contiene* un'altra foresta F' se F contiene F' come grafo diretto. Chiaramente, se F contiene come sottoinsieme proprio F', allora F ha meno componenti di F'. La figura mostra due di tali foreste con le loro radici in cima.



F_2 contiene F_3



Ecco l'idea cruciale. Si chiami una successione $F_1, \ldots, F_k$ di foreste una *successione raffinante* se $F_i \in \mathcal{F}_{n,i}$ e F_i contiene F_{i+1}, per ogni i.

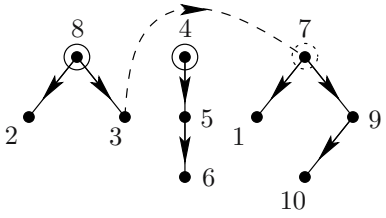
Ora sia F_k una foresta fissata in $\mathcal{F}_{n,k}$ e si indichi

- con $N(F_k)$ il numero di alberi con radice contententi F_k, e
- con $N^*(F_k)$ il numero di successioni raffinanti che terminano in F_k.

Contiamo $N^*(F_k)$ in due modi, prima iniziando da un albero e quindi iniziando da F_k. Si supponga che $F_1 \in \mathcal{F}_{n,1}$ contenga F_k. Poiché possiamo eliminare i $k-1$ archi di $F_1 \backslash F_k$ in ogni ordine possibile per ottenere una successione raffinante da F_1 a F_k, troviamo

$$N^*(F_k) \;=\; N(F_k)\,(k-1)!. \tag{3}$$

Cominciamo ora dall'altro estremo. Per produrre da F_k un F_{k-1} dobbiamo aggiungere l'arco orientato, da ogni vertice a, ad ognuna delle $k-1$



$F_3 \longrightarrow F_2$

radici degli alberi che non contengono a (si veda la figura a margine, dove passiamo da F_3 a F_2 aggiungendo un arco 3●➔●7). Pertanto abbiamo $n(k-1)$ scelte. Analogamente, per F_{k-1} possiamo produrre un arco orientato da ogni vertice b ad ognuna delle $k-2$ radici degli alberi non contenenti b. Per questo abbiamo $n(k-2)$ scelte. Continuando in questo modo, giungiamo a

$$N^*(F_k) \;=\; n^{k-1}(k-1)!, \tag{4}$$

ed ecco, con (3), la relazione di inaspettata semplicità

$$N(F_k) = n^{k-1} \qquad \text{per } \textit{ogni } F_k \in \mathcal{F}_{n,k}.$$

Per $k = n$, F_n consiste solo in n vertici isolati. Pertanto $N(F_n)$ conta il numero di *tutti* gli alberi con radice ed otteniamo $|\mathcal{F}_{n,1}| = n^{n-1}$ e dunque la formula di Cayley. □

Ma possiamo ottenere anche di più da questa dimostrazione. La formula (4) dà per $k = n$:

$$\#\big\{\text{successioni raffinanti } (F_1, F_2, \ldots, F_n)\big\} \;=\; n^{n-1}(n-1)!. \tag{5}$$

Per $F_k \in \mathcal{F}_{n,k}$, si indichi con $N^{**}(F_k)$ il numero delle successioni raffinanti $F_1, \ldots, F_n$ il cui termine k-esimo è F_k. Chiaramente questo è $N^*(F_k)$ volte il numero dei modi di scegliere $(F_{k+1}, \ldots, F_n)$. Ma quest'ultimo numero è $(n-k)!$ poiché possiamo eliminare gli $n-k$ archi di F_k in ogni modo possibile, e dunque

$$N^{**}(F_k) \;=\; N^*(F_k)(n-k)! \;=\; n^{k-1}(k-1)!(n-k)!. \tag{6}$$

Poiché tale numero non dipende dalla scelta di F_k, dividendo (5) per (6) si ottiene il numero di foreste radicate con k alberi:

$$|\mathcal{F}_{n,k}| \;=\; \frac{n^{n-1}(n-1)!}{n^{k-1}(k-1)!(n-k)!} \;=\; \binom{n}{k} k\, n^{n-1-k}.$$

Poiché possiamo scegliere le k radici in $\binom{n}{k}$ modi possibili, abbiamo ridimostrato la formula $T_{n,k} = kn^{n-k-1}$ senza ricorrere all'induzione.

Terminiamo con una nota storica. La pubblicazione di Cayley del 1889 fu anticipata da Carl W. Borchardt (1860) e questo fatto fu riconosciuto dallo stesso Cayley. Un risultato equivalente apparve addirittura precedentemente in una pubblicazione di James J. Sylvester (1857), si veda il Capitolo 3 in [2]. La novità nella pubblicazione di Cayley fu l'uso della terminologia della teoria dei grafi ed in seguito il teorema è sempre stato associato al suo nome.

Bibliografia

[1] M. AIGNER: *Combinatorial Theory,* Springer-Verlag, Berlin Heidelberg New York 1979; Reprint 1997.

[2] N. L. BIGGS, E. K. LLOYD & R. J. WILSON: *Graph Theory 1736-1936,* Clarendon Press, Oxford 1976.

[3] A. CAYLEY: *A theorem on trees,* Quart. J. Pure Appl. Math. **23** (1889), 376-378; Collected Mathematical Papers Vol. 13, Cambridge University Press 1897, 26-28.

[4] A. JOYAL: *Une théorie combinatoire des séries formelles,* Advances in Math. **42** (1981), 1-82.

[5] J. PITMAN: *Coalescent random forests,* J. Combinatorial Theory, Ser. A **85** (1999), 165-193.

[6] H. PRÜFER: *Neuer Beweis eines Satzes über Permutationen,* Archiv der Math. u. Physik (3) **27** (1918), 142-144.

[7] A. RÉNYI: *Some remarks on the theory of trees.* MTA Mat. Kut. Inst. Kozl. (Publ. math. Inst. Hungar. Acad. Sci.) **4** (1959), 73-85; Selected Papers Vol. 2, Akadémiai Kiadó, Budapest 1976, 363-374.

[8] J. RIORDAN: *Forests of labeled trees,* J. Combinatorial Theory **5** (1968), 90-103.

Completando i quadrati latini

Capitolo 27

Alcuni dei più vecchi oggetti del calcolo combinatorio, il cui studio apparentemente risale ai tempi antichi, sono i *quadrati latini*. Per ottenere un quadrato latino, bisogna riempire le n^2 caselle di una matrice quadrata $(n \times n)$ con i numeri $1, 2, \ldots, n$ in modo tale che ogni numero appaia esattamente una volta in ogni riga ed in ogni colonna. In altre parole, ogni riga ed ogni colonna rappresentano una permutazione dell'insieme $\{1, \ldots, n\}$. Chiamiamo n l'*ordine* del quadrato latino.

1	2	3	4
2	1	4	3
4	3	1	2
3	4	2	1

Un quadrato latino di ordine 4

Ecco il problema che vogliamo discutere. Supponiamo che qualcuno abbia iniziato a riempire le caselle con i numeri $\{1, 2, \ldots, n\}$. Ad un certo punto smette e ci chiede di riempire le caselle rimanenti in modo da ottenere un quadrato latino. Quando è possibile farlo? Per avere una possibilità dobbiamo, ovviamente, supporre che all'inizio del nostro compito ogni elemento appaia al massimo una volta in ogni riga ed in ogni colonna. Diamo un nome a questa situazione. Parliamo di *quadrato latino parziale* di ordine n se alcune caselle di una matrice $(n \times n)$ sono riempite con numeri dall'insieme $\{1, \ldots, n\}$ in modo tale che ogni numero appaia al massimo una volta in ogni riga e colonna. Pertanto il problema è:

Quando su può completare un quadrato latino parziale per formare un quadrato latino dello stesso ordine?

1	4	2	5	3
4	2	5	3	1
2	5	3	1	4
5	3	1	4	2
3	1	4	2	5

Un quadrato latino ciclico

Consideriamo qualche esempio. Supponiamo che le prime $n-1$ righe siano riempite e che l'ultima riga sia vuota. Allora possiamo facilmente riempire l'ultima riga. Si noti che ogni elemento appare $n - 1$ volte nel quadrato latino parziale e pertanto manca da esattamente una colonna. Dunque scrivendo ogni elemento sotto la colonna da cui manca abbiamo completato il quadrato correttamente.

Andando all'altro estremo, supponiamo che sia riempita solo la prima riga. Allora è ancora facile completare il quadrato ruotando ciclicamente gli elementi di un passo in ognuna delle righe seguenti.

Pertanto, mentre nel nostro primo esempio il completamento è forzato, abbiamo molte possibilità nel secondo esempio. In generale, meno caselle sono riempite, più libertà dovremmo avere nel completare il quadrato.

Tuttavia, a margine si mostra un esempio di un quadrato parziale con solo n caselle riempite che chiaramente non può essere completato, poiché non vi è modo di riempire l'angolo in alto a destra senza violare la condizione sulle righe o quella sulle colonne.

1	2	...	n-1	
				n

Un quadrato latino parziale che non può essere completato

Se sono riempite meno di n caselle in una matrice $(n \times n)$, è sempre possibile completare la matrice in modo da ottenere un quadrato latino?

Questa domanda fu sollevata da Trevor Evans nel 1960 e l'affermazione che un completamento è sempre possibile divenne rapidamente nota come la congettura di Evans. Ovviamente, si potrebbe tentare con l'induzione, e questo è ciò che alla fine portò al successo. Ma la dimostrazione di Bohdan Smetaniuk del 1981, che rispose alla domanda, è uno splendido esempio di quanto possa essere sottile una dimostrazione per induzione necessaria per un tale compito. Non solo, la dimostrazione è costruttiva, in quanto ci mostra come completare il quadrato latino esplicitamente a partire da ogni configurazione parziale iniziale.

Prima di procedere alla dimostrazione diamo un'occhiata più ravvicinata ai quadrati latini in generale. Possiamo alternativamente vedere un quadrato latino come array $(3 \times n^2)$, chiamato l'*array delle linee* del quadrato latino. La figura a margine mostra un quadrato latino di ordine 3 e l'array delle linee ad esso associato, dove R, C e E stanno per righe, colonne ed elementi.

1	3	2
2	1	3
3	2	1

R: 1 1 1 2 2 2 3 3 3
C: 1 2 3 1 2 3 1 2 3
E: 1 3 2 2 1 3 3 2 1

La condizione sul quadrato latino è equivalente a dire che, in ogni coppia di linee dell'array delle linee, appaiono tutte le n^2 coppie ordinate (e pertanto ogni coppia appare esattamente una volta). Chiaramente, possiamo permutare i simboli in ogni linea arbitrariamente (il che corrisponde a permutazioni di righe, colonne ed elementi) ed ottenere ancora un quadrato latino. Ma la condizione sull'array $(3 \times n^2)$ ci dice di più: non vi è un ruolo speciale per gli elementi. Possiamo anche permutare le linee dell'intero array e conservare ancora le condizioni sull'array, ottenendo pertanto un quadrato latino.

I quadrati latini collegati da ognuna di tali permutazioni sono definiti *coniugati*. Ecco l'osservazione che renderà trasparente la dimostrazione: un quadrato latino parziale corrisponde ovviamente ad un array delle linee parziale (ogni coppia appare al massimo una volta in ogni coppia di linee), ed ogni coniugato di un quadrato latino parziale è anch'esso un quadrato latino parziale. In particolare, un quadrato latino parziale può essere completato se e solo se ogni coniugato può essere completato (basta completare un coniugato e quindi invertire la permutazione delle tre linee).

Se permutiamo le righe dell'esempio riportato sopra ciclicamente, $R \longrightarrow C \longrightarrow E \longrightarrow R$, allora otteniamo l'array delle linee ed il quadrato latino seguenti:

1	2	3
3	1	2
2	3	1

R : 1 3 2 2 1 3 3 2 1
C: 1 1 1 2 2 2 3 3 3
E: 1 2 3 1 2 3 1 2 3

Avremo bisogno di due risultati, dovuti a Herbert J. Ryser ed a Charles C. Lindner, noti prima del teorema di Smetaniuk. Se un quadrato latino parziale è di forma tale che le prime r righe sono completamente riempite e le restanti caselle sono vuote, allora parliamo di un *rettangolo latino* $(r \times n)$.

Lemma 1. *Ogni rettangolo latino $(r \times n)$ $r < n$, può essere esteso ad un rettangolo latino $((r+1) \times n)$ e pertanto può essere completato come quadrato latino.*

■ **Dimostrazione.** Applichiamo il teorema di Hall (si veda al Capitolo 23).

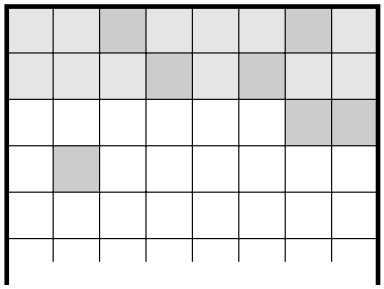
Sia A_j l'insieme dei numeri che *non* appaiono nella colonna j. Una riga $(r+1)$-esima ammissibile corrisponde allora precisamente a un sistema di rappresentanti distinti per la collezione $A_1, \ldots, A_n$. Per dimostrare il lemma dobbiamo pertanto verificare la condizione di Hall (H). Ogni insieme A_j ha dimensione $n-r$, ed ogni elemento è precisamente in $n-r$ insiemi A_j (poiché appare r volte nel rettangolo). Ogni gruppo di m fra gli insiemi A_j contiene globalmente $m(n-r)$ elementi e pertanto almeno m elementi diversi, il che è esattamente la condizione (H). □

Lemma 2. *Sia P un quadrato latino parziale di ordine n con al massimo $n-1$ caselle riempite ed al massimo $\frac{n}{2}$ elementi distinti; allora P può essere completato come quadrato latino di ordine n.*

■ **Dimostrazione.** Cominciamo trasformando il problema in una forma più conveniente.

In base al principio di coniugazione discusso prima, possiamo sostituire la condizione "al massimo $\frac{n}{2}$ elementi distinti" con la condizione che gli elementi appaiano in al più $\frac{n}{2}$ righe e possiamo inoltre supporre che queste righe siano le righe in alto. Dunque supponiamo che le righe con le caselle riempite siano quelle in posizione $1, 2, \ldots, r$, con f_i caselle riempite nella riga i, dove $r \leq \frac{n}{2}$ e $\sum_{i=1}^{r} f_i \leq n-1$. Permutando le righe, possiamo supporre che $f_1 \geq f_2 \geq \ldots \geq f_r$. Ora completiamo le righe $1, \ldots, r$ passo dopo passo fino a raggiungere un rettangolo $(r \times n)$ che può quindi essere esteso ad un quadrato latino in base al Lemma 1.

Supponiamo di aver già riempito le righe $1, 2, \ldots, \ell-1$. Nella riga ℓ vi sono f_ℓ caselle riempite che possiamo supporre essere quelle in fondo. La situazione corrente è mostrata nella figura, dove la parte ombreggiata indica le caselle riempite.



Una situazione per $n = 8$, con $\ell = 3$, $f_1 = f_2 = f_3 = 2$, $f_4 = 1$. I quadrati scuri rappresentano le caselle pre-riempite, quelli più chiari mostrano le caselle riempite nel processo di completamento

Il completamento della riga ℓ si svolge mediante un'altra applicazione del teorema di Hall, ma questa volta è alquanto sottile. Sia X l'insieme degli elementi che *non* appaiono nella riga ℓ, pertanto $|X| = n - f_\ell$, e per $j = 1, \ldots, n - f_\ell$ sia A_j l'insieme degli elementi di X che *non* appaiono nella colonna j (né sopra né sotto la riga ℓ). Per completare la riga ℓ dobbiamo quindi verificare la condizione (H) per la collezione $A_1, \ldots, A_{n-f_\ell}$.

In primo luogo affermiamo

$$n - f_\ell - \ell + 1 \;>\; \ell - 1 + f_{\ell+1} + \ldots + f_r. \tag{1}$$

Il caso $\ell = 1$ è chiaro. Altrimenti $\sum_{i=1}^{r} f_i < n$, $f_1 \geq \ldots \geq f_r$ e $1 < \ell \leq r$ insieme implicano

$$n \;>\; \sum_{i=1}^{r} f_i \;\geq\; (\ell-1) f_{\ell-1} + f_\ell + \ldots + f_r.$$

Ora o $f_{\ell-1} \geq 2$ (in tal caso vale (1)) oppure $f_{\ell-1} = 1$. Nel secondo caso, (1) si riduce a $n > 2(\ell-1) + r - \ell + 1 = r + \ell - 1$, il che è vero poiché $\ell \leq r \leq \frac{n}{2}$.

Consideriamo ora m insiemi A_j, $1 \leq m \leq n - f_\ell$, e sia B la loro unione. Dobbiamo mostrare che $|B| \geq m$. Si consideri il numero c di caselle nelle m colonne corrispondenti agli A_j che contengono elementi di X. Vi sono al massimo $(\ell-1)m$ di tali caselle sopra la riga ℓ ed al massimo $f_{\ell+1}+\ldots+f_r$ al di sotto della riga ℓ, pertanto

$$c \;\leq\; (\ell-1)m + f_{\ell+1} + \ldots + f_r.$$

D'altro canto, ogni elemento $x \in X\backslash B$ appare in ognuna delle m colonne, dunque $c \geq m(|X| - |B|)$ e pertanto (con $|X| = n - f_\ell$)

$$|B| \;\geq\; |X| - \tfrac{1}{m}c \;\geq\; n - f_\ell - (\ell-1) - \tfrac{1}{m}(f_{\ell+1} + \ldots + f_r).$$

Ne segue che $|B| \geq m$ se

$$n - f_\ell - (\ell-1) - \tfrac{1}{m}(f_{\ell+1} + \ldots + f_r) \;>\; m - 1,$$

ovvero, se

$$m(n - f_\ell - \ell + 2 - m) \;>\; f_{\ell+1} + \ldots + f_r. \tag{2}$$

La disuguaglianza (2) è vera per $m = 1$ e per $m = n - f_\ell - \ell + 1$ in base a (1) e dunque per tutti i valori m tra 1 e $n - f_\ell - \ell + 1$, poiché il membro di sinistra è una funzione quadratica in m con coefficiente principale -1. Il caso rimanente è $m > n - f_\ell - \ell + 1$. Poiché ogni elemento x di X è contenuto in al più $\ell - 1 + f_{\ell+1} + \ldots + f_r$ righe, esso può anche apparire al massimo nello stesso numero di colonne. Ricorrendo ancora una volta a (1), troviamo che x è in uno degli insiemi A_j, così in questo caso $B = X$, $|B| = n - f_\ell \geq m$, e la dimostrazione è completa. □

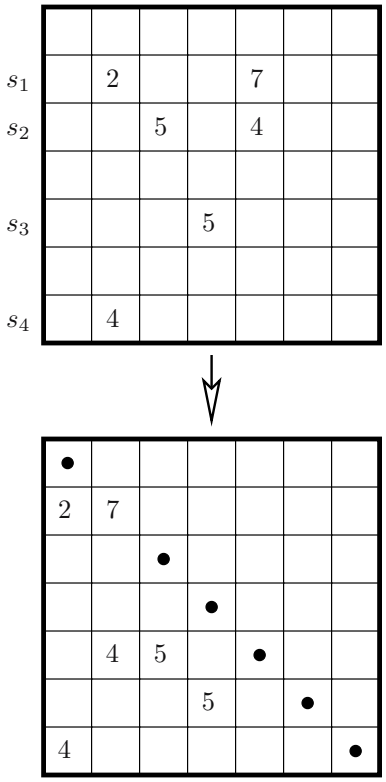
Dimostriamo infine il teorema di Smetaniuk.

Teorema. *Ogni quadrato latino parziale di ordine n con al massimo $n-1$ caselle riempite può essere completato come un quadrato latino dello stesso ordine.*

■ **Dimostrazione.** Usiamo l'induzione su n, i casi $n \leq 2$ essendo ovvii. Studiamo pertanto ora un quadrato latino parziale di ordine $n \geq 3$ con al massimo $n-1$ caselle riempite. Con la notazione introdotta sopra tali caselle si trovano in $r \leq n-1$ righe differenti numerate $s_1, \ldots, s_r$, che contengono $f_1, \ldots, f_r > 0$ caselle riempite, con $\sum_{i=1}^{r} f_i \leq n-1$. In base al Lemma 2 possiamo supporre che vi siano più di $\frac{n}{2}$ elementi diversi; pertanto vi è un elemento che appare una sola volta: dopo aver rinumerato e permutato le righe (se necessario) possiamo supporre che l'elemento n appaia una sola volta, e questo nella riga s_1.

Al passo successivo vogliamo permutare le righe e le colonne del quadrato latino parziale in modo tale che dopo le permutazioni tutte le caselle riempite giacciano sotto la diagonale, ad eccezione della casella riempita con n, che finirà sulla diagonale. (La diagonale consiste nelle caselle (k,k) tali che $1 \leq k \leq n$.) Per ottenerla, permutiamo dapprima la riga s_1 in posizione

f_1. Mediante permutazione delle colonne spostiamo tutte le caselle riempite a sinistra, cosicché n appaia come ultimo elemento nella sua riga, sulla diagonale. Quindi spostiamo la riga s_2 in posizione $1 + f_1 + f_2$ e di nuovo le caselle riempite il più a sinistra possibile. In generale, per $1 < i \leq r$ spostiamo la riga s_i in posizione $1 + f_1 + f_2 + \ldots + f_i$ e le caselle riempite il più a sinistra possibile. Ciò dà chiaramente la configurazione desiderata. Il disegno a margine mostra un esempio, con $n = 7$: le righe $s_1 = 2$, $s_2 = 3$, $s_3 = 5$ e $s_4 = 7$ con $f_1 = f_2 = 2$ e $f_3 = f_4 = 1$ sono spostate nelle righe numerate 2, 5, 6 e 7 e le colonne sono permutate a sinistra in modo che alla fine tutti gli elementi eccetto il 7 vengano a trovarsi sotto la diagonale, che è indicata con i simboli •.



Al fine di poter applicare l'induzione eliminiamo ora l'elemento n dalla diagonale ed ignoriamo la prima riga e l'ultima colonna (che non contengono alcuna casella riempita): pertanto stiamo considerando un quadrato latino parziale di ordine $n - 1$ con al più $n - 2$ caselle riempite, che per induzione possono essere completate come un quadrato latino di ordine $n - 1$. La figura a margine mostra uno dei (molti) completamenti del quadrato latino parziale che sorge dal nostro esempio. Nella figura, gli elementi originali sono stampati in grassetto. Essi sono già definitivi, come anche tutti gli elementi nelle caselle ombreggiate; alcuni degli altri elementi saranno cambiati in seguito, in modo da completare il quadrato latino di ordine n.

Al passo successivo vogliamo spostare gli elementi diagonali del quadrato nell'ultima colonna e mettere gli elementi n sulla diagonale al loro posto. Tuttavia, in generale non potremo farlo in quanto gli elementi non saranno necessariamente distinti. Pertanto procediamo più attentamente ed effettuiamo successivamente, per $k = 2, 3, \ldots, n - 1$ (in quest'ordine), la seguente operazione:

Si metta il valore n nella casella (k, n). In tal modo si trova un quadrato latino parziale corretto. Ora si scambi il valore x_k nella casella diagonale (k, k) con il valore n nella casella (k, n) nell'ultima colonna.

Se il valore x_k non è già apparso nell'ultima colonna, allora il nostro compito per il k corrente è completato. Dopo questo, gli elementi correnti nella colonna k-esima non saranno più cambiati.

Nel nostro esempio questo funziona senza problemi per $k = 2$, 3 e 4, e gli elementi diagonali corrispondenti 3, 1 e 6 si spostano nell'ultima colonna.

2	3	4	1	6	5	
5	6	1	4	2	3	
1	2	3	6	5	4	
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

Le tre figure seguenti mostrano le operazioni corrispondenti.

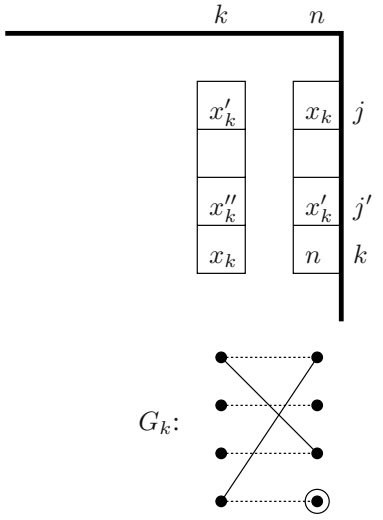
2	7	4	1	6	5	3
5	6	7	4	2	3	1
1	2	3	6	5	4	7
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

2	3	4	1	6	5	7
5	6	1	4	2	3	
1	2	3	6	5	4	
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

2	7	4	1	6	5	3
5	6	1	4	2	3	7
1	2	3	6	5	4	
6	**4**	**5**	2	3	1	
3	1	6	**5**	4	2	
4	5	2	3	1	6	

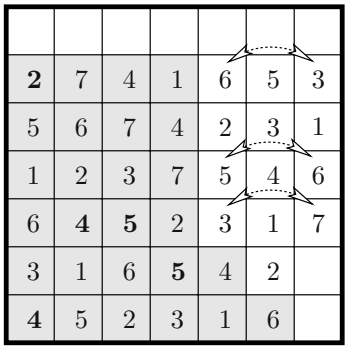
Dobbiamo ora trattare il caso in cui vi sia già un elemento x_k nell'ultima colonna. In tal caso procediamo come segue:

Se vi è già un elemento x_k in una casella (j, n) tale che $2 \leq j < k$, allora scambiamo nella riga j l'elemento x_k della n-esima colonna con l'elemento x'_k della k-esima colonna. Se l'elemento x'_k appare anche in una casella (j', n), allora scambiamo anche gli elementi nella riga j'-esima che appaiono nelle colonne n-esima e k-esima, e così via.

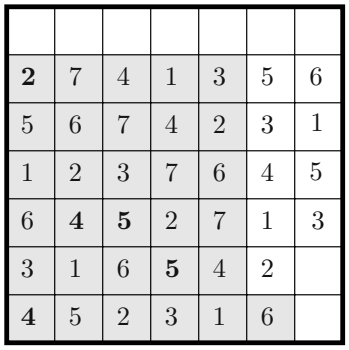


Se procediamo in questo modo non ci saranno mai due elementi uguali in una riga. Il nostro procedimento di scambio assicura che non vi saranno neppure due elementi uguali in una colonna. Pertanto dobbiamo solo controllare che il processo di scambio tra la k-esima e la n-esima colonna non porti ad un ciclo infinito. Ciò si può verificare osservando il seguente grafo bipartito G_k: i suoi vertici corrispondono alle caselle (i, k) e (j, n) con $2 \leq i, j \leq k$ i cui elementi possono essere scambiati. Vi è uno spigolo tra (i, k) e (j, n) se queste due caselle giacciono nella stessa riga (ovvero, con $i = j$) oppure se le caselle prima del processo di scambio contengono lo stesso elemento (il che implica $i \neq j$). Nel nostro schizzo gli spigoli con $i = j$ sono punteggiati, gli altri no. Tutti i vertici in G_k hanno grado 1 o 2. La casella (k, n) corrisponde ad un vertice di grado 1; tale vertice è l'inizio di un cammino che porta alla colonna k su uno spigolo orizzontale, quindi possibilmente su uno spigolo diagonale di nuovo alla colonna n, quindi orizzontalmente ancora alla colonna k e così via. Il processo termina nella colonna k con un valore che non appare nella colonna n. Pertanto le operazioni di scambio termineranno ad un certo punto con un passo in cui spostiamo un *nuovo* elemento nell'ultima colonna. Quindi il lavoro sulla colonna k è completo e gli elementi nelle caselle (i, k) per $i \geq 2$ sono definitivamente fissati.

Nel nostro esempio il "caso dello scambio" si verifica per $k = 5$: l'elemento $x_5 = 3$ appare già nell'ultima colonna, cosicché l'elemento deve essere spostato indietro nella colonna $k = 5$. Ma nemmeno l'elemento di scambio $x'_5 = 6$ è nuovo; esso viene scambiato con $x''_5 = 5$, e quest'ultimo è nuovo.

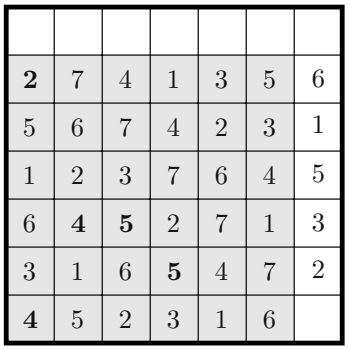


2	7	4	1	6	5	3
5	6	7	4	2	3	1
1	2	3	7	5	4	6
6	**4**	**5**	2	3	1	7
3	1	6	**5**	4	2	
4	5	2	3	1	6	

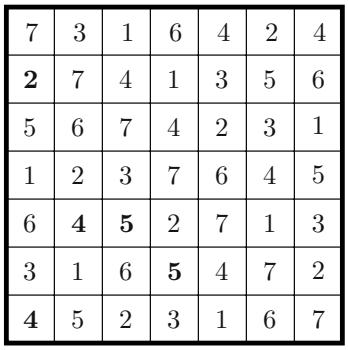


2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	2	
4	5	2	3	1	6	

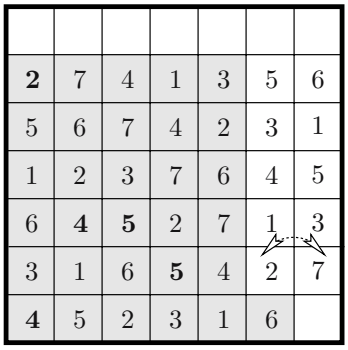
Infine lo scambio per $k = 6 = n - 1$ non pone alcun problema, e dopo questo il completamento del quadrato latino è unico:



2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	7	2
4	5	2	3	1	6	



7	3	1	6	4	2	4
2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	7	2
4	5	2	3	1	6	7



2	7	4	1	3	5	6
5	6	7	4	2	3	1
1	2	3	7	6	4	5
6	**4**	**5**	2	7	1	3
3	1	6	**5**	4	2	7
4	5	2	3	1	6	

...e lo stesso avviene in generale: poniamo un elemento n nella casella (n, n) e dopo questo la prima riga può essere completata con gli elementi mancanti delle rispettive colonne (si veda il Lemma 1); questo completa la dimostrazione. Al fine di ottenere esplicitamente un completamento del quadrato latino parziale originale di ordine n, ci basta invertire le permutazioni degli elementi, delle righe e delle colonne dei primi due passi della dimostrazione. □

Bibliografia

[1] T. EVANS: *Embedding incomplete Latin squares,* Amer. Math. Monthly **67** (1960), 958-961.

[2] C. C. LINDNER: *On completing Latin rectangles,* Canadian Math. Bulletin **13** (1970), 65-68.

[3] H. J. RYSER: *A combinatorial theorem with an application to Latin rectangles,* Proc. Amer. Math. Soc. **2** (1951), 550-552.

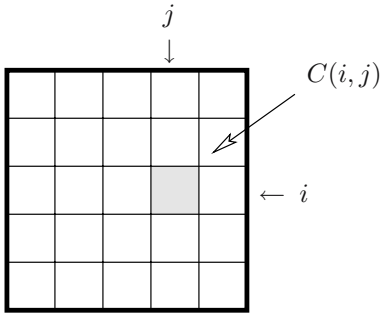
[4] B. SMETANIUK: *A new construction on Latin squares I: A proof of the Evans conjecture,* Ars Combinatoria **11** (1981), 155-172.

Il problema di Dinitz

Capitolo 28

Il problema dei quattro colori ha costituito una delle principali forze trainanti per lo sviluppo della teoria dei grafi come la conosciamo oggi e la colorazione è tuttora uno degli argomenti preferiti dai teorici dei grafi. Ecco un problema di colorazione all'apparenza semplice, sollevato da Jeff Dinitz nel 1978, il quale resistette a tutti gli attacchi fino alla sua soluzione di una semplicità sorprendente data da Fred Galvin quindici anni più tardi.

> *Si considerino n^2 caselle disposte in un quadrato $(n \times n)$, e sia (i, j) la casella nella riga i e nella colonna j. Si supponga che per ogni casella (i, j) abbiamo un insieme $C(i, j)$ di n colori.*
> *È allora sempre possibile colorare l'intera matrice scegliendo per ogni casella (i, j) un colore dal proprio insieme $C(i, j)$ tale che i colori in ogni riga ed in ogni colonna siano distinti?*



Per cominciare si consideri il caso in cui tutti gli insiemi di colori $C(i, j)$ siano uguali, diciamo $\{1, 2, \ldots, n\}$. Allora il problema di Dinitz si riduce al compito seguente: riempire il quadrato $(n \times n)$ con i numeri $1, 2, \ldots, n$ in modo tale che i numeri in ogni riga e colonna siano distinti. In altre parole, ognuna di tali colorazioni corrisponde ad un quadrato latino, come discusso nel capitolo precedente. Pertanto, in questo caso, la risposta alla nostra domanda è "sì ".

Dal momento che questo è tanto semplice, perché dovrebbe essere molto più difficile il caso generale in cui l'insieme $C := \bigcup_{i,j} C(i, j)$ contiene anche più di n colori? La difficoltà deriva dal fatto che non ogni colore di C è disponibile in ogni casella. Ad esempio, mentre nel caso dei quadrati latini possiamo chiaramente scegliere una permutazione arbitraria dei colori nella prima riga, questo non è piu valido nel problema generale. Tale difficoltà è illustrata già dal caso $n = 2$.

$\{1,2\}$	$\{2,3\}$
$\{1,3\}$	$\{2,3\}$

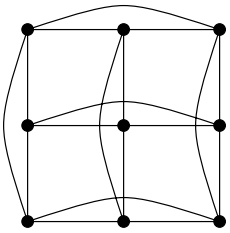
Supponiamo di avere gli insiemi di colori indicati nella figura. Se scegliamo i colori 1 e 2 dalla prima riga, allora siamo in difficoltà poiché dovremmo scegliere il colore 3 per entrambe le caselle nella seconda colonna.

Prima di affrontare il problema di Dinitz, riformuliamo la situazione nel linguaggio della teoria dei grafi. Come al solito consideriamo solo grafi $G = (V, E)$ senza cicli e spigoli multipli. Si denoti con $\chi(G)$ il *numero cromatico* del grafo ovvero il più piccolo numero di colori che si possano assegnare ai vertici in modo tale che i vertici adiacenti ricevano colori diversi.

In altre parole, una colorazione si riduce ad una partizione di V in classi (colorate con lo stesso colore) tale che non vi siano spigoli in una classe. Definendo un insieme $A \subseteq V$ *indipendente* se non vi sono spigoli in A, deduciamo che il numero cromatico è il più piccolo numero di insiemi indipendenti che partizionano l'insieme dei vertici V.

Nel 1976 Vizing, e tre anni più tardi Erdős, Rubin e Taylor, studiarono la seguente variante della colorazione che ci porta direttamente al problema di Dinitz. Supponiamo che nel grafo $G = (V, E)$ abbiamo un insieme $C(v)$ di colori per ogni vertice v. Una *colorazione per lista* è una colorazione $c : V \longrightarrow \bigcup_{v \in V} C(v)$ tale che $c(v) \in C(v)$ per ogni $v \in V$. La definizione di *numero cromatico della lista* $\chi_\ell(G)$ dovrebbe ora essere chiara: è il più piccolo numero k tale per cui *ogni* lista di insiemi di colori $C(v)$ con $|C(v)| = k$ per ogni $v \in V$ vi sia sempre una colorazione per lista. Ovviamente, abbiamo $\chi_\ell(G) \leq |V|$ (non esauriamo mai i colori). Poiché la colorazione ordinaria è semplicemente il caso speciale di colorazione della lista in cui tutti gli insiemi $C(v)$ sono uguali, otteniamo per ogni grafo G

$$\chi(G) \ \leq \ \chi_\ell(G).$$



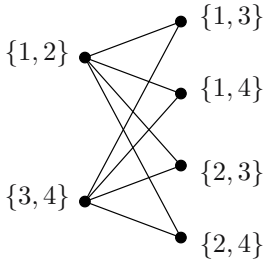
Il grafo S_3

Per tornare al problema di Dinitz, si consideri il grafo S_n che ha come insieme di vertici le n^2 caselle della nostra matrice $(n \times n)$ dove due caselle sono adiacenti se e solo se sono nella stessa riga o colonna.

Poiché in ogni gruppo di n caselle in una riga queste sono adiacenti a due a due, abbiamo bisogno di almeno n colori. Inoltre, ogni colorazione con n colori corrisponde ad un quadrato latino, in cui le caselle occupate dallo stesso numero formano una classe di colori. Poiché i quadrati latini, come abbiamo visto, esistono, deduciamo che $\chi(S_n) = n$ ed il problema di Dinitz può ora essere succintamente espresso come

$$\chi_\ell(S_n) \ = \ n?$$

Si potrebbe ipotizzare che per ogni grafo G valga $\chi(G) = \chi_\ell(G)$, ma questo è molto distante dal vero. Si consideri il grafo $G = K_{2,4}$. Il numero cromatico è 2 poiché possiamo usare un colore per i due vertici a sinistra ed il secondo per i vertici a destra. Ma supponiamo ora di avere gli insiemi di colori indicati nella figura.



Per colorare i vertici di sinistra abbiamo le quattro possibilità $1|3$, $1|4$, $2|3$ e $2|4$, ma ognuna di queste coppie appare come un colore assegnato a destra, pertanto una colorazione per lista non è possibile. Quindi $\chi_\ell(G) \geq 3$ e il lettore potrà trovare divertente dimostrare che $\chi_\ell(G) = 3$ (non c'è bisogno di provare tutte le possibilità!). Generalizzando questo esempio, non è difficile trovare grafi G in cui $\chi(G) = 2$, ma $\chi_\ell(G)$ sia arbitrariamente grande! Pertanto il problema della colorazione per lista non è semplice come sembra al primo sguardo.

Torniamo al problema di Dinitz. Un passo significativo verso la soluzione fu compiuto da Jeanette Janssen nel 1992 quando dimostrò che $\chi_\ell(S_n) \leq n + 1$ ed il *coup de grâce* fu dato da Fred Galvin che combinò ingegnosamente due risultati, entrambi noti da lungo tempo. Discuteremo questi due risultati e mostreremo quindi come essi implichino $\chi_\ell(S_n) = n$.

Fissiamo dapprima alcune notazioni. Supponiamo che v sia un vertice del grafo G, allora come prima indichiamo con $d(v)$ il *grado* di v. Nel nostro grafo quadrato S_n ogni vertice ha grado $2n-2$, tenendo conto degli altri $n-1$ vertici nella stessa riga e nella stessa colonna. Per un sottoinsieme $A \subseteq V$ indichiamo con G_A il sottografo che ha come insieme di vertici A e che contiene tutti gli spigoli di G tra i vertici di A. Chiamiamo G_A il sottografo indotto da A e diciamo che H è un *sottografo indotto* di G se $H = G_A$ per un dato A.

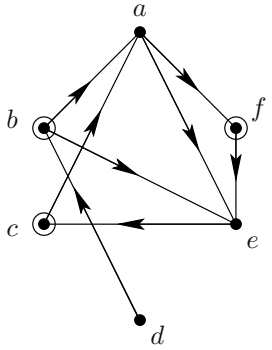
Per enunciare il nostro primo risultato abbiamo bisogno di *grafi orientati* $\vec{G} = (V, E)$ ovvero grafi in cui ogni spigolo e ha un orientamento. La notazione $e = (u, v)$ significa che vi è un arco e, indicato anche come $u \longrightarrow v$, il cui vertice iniziale è u e il cui vertice finale è v. Ha allora senso parlare del *grado esterno* $d^+(v)$ risp. del *grado interno* $d^-(v)$, dove $d^+(v)$ conta il numero di spigoli con v come vertice iniziale ed analogamente per $d^-(v)$; inoltre, $d^+(v) + d^-(v) = d(v)$. Quando scriviamo G, intendiamo il grafo $\vec{G}$ senza gli orientamenti.

Il seguente concetto ebbe origine nell'analisi dei giochi e avrà un ruolo cruciale nella nostra discussione.

Definizione 1. Sia $\vec{G} = (V, E)$ un grafo orientato. Un *nucleo* $K \subseteq V$ è un sottoinsieme dei vertici tale che

(i) K è indipendente in G, e

(ii) per ogni $u \notin K$ esiste un vertice $v \in K$ con uno spigolo $u \longrightarrow v$.



Analizziamo l'esempio nella figura. L'insieme $\{b, c, f\}$ costituisce un nucleo, ma il sottografo indotto da $\{a, c, e\}$ non ha un kernel poiché i tre spigoli si alternano attraverso i vertici.

Forti di tutte queste premesse siamo ora pronti per affermare il primo risultato.

Lemma 1. *Sia $\vec{G} = (V, E)$ un grafo orientato e si supponga che per ogni vertice $v \in V$ abbiamo un insieme di colori $C(v)$ più grande del grado esterno, $|C(v)| \geq d^+(v) + 1$. Se ogni sottografo indotto di $\vec{G}$ possiede un nucleo, allora esiste una colorazione per lista G con un colore da $C(v)$ per ogni v.*

■ **Dimostrazione.** Procediamo per induzione su $|V|$. Per $|V| = 1$ non vi è nulla da dimostrare. Si scelga un colore $c \in C = \bigcup_{v \in V} C(v)$ e si ponga

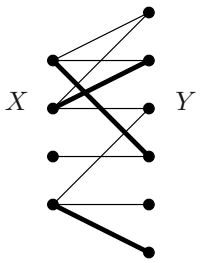
$$A(c) := \{v \in V : c \in C(v)\}.$$

Per ipotesi, il sottografo indotto $G_{A(c)}$ possiede un nucleo $K(c)$. Ora coloriamo tutti i $v \in K(c)$ con il colore c (ciò è possibile poiché $K(c)$ è indipendente) ed eliminiamo $K(c)$ da G e c da C. Sia G' il sottografo indotto di G su $V \backslash K(c)$ con $C'(v) = C(v) \backslash c$ come nuova lista degli insiemi di colori. Si noti che per ogni $v \in A(c) \backslash K(c)$, il grado esterno $d^+(v)$ è ridotto di almeno 1 (a causa della condizione (ii) nelle definizione di nucleo). Pertanto $d^+(v) + 1 \leq |C'(v)|$ vale ancora in $\vec{G}'$. La stessa condizione vale

anche per i vertici fuori da $A(c)$, poiché in questo caso gli insiemi di colori $C(v)$ restano invariati. Il nuovo grafo G' contiene meno vertici di G e possiamo concludere con un argomento di induzione. □

Il metodo di attacco per il problema di Dinitz è ora ovvio: dobbiamo trovare un orientamento del grafo S_n che abbia grado più alto di $d^+(v) \leq n-1$ per tutti i vertici v e che assicuri l'esistenza di un kernel per tutti i sottografi indotti. Ciò è compiuto dal nostro secondo risultato.

Ci serve ancora qualche preparazione. Si ricordi (dal Capitolo 9) che un grafo *bipartito* $G = (X \cup Y, E)$ è un grafo con la proprietà seguente: l'insieme dei vertici V è diviso in due parti X e Y in modo tale che ogni spigolo abbia un estremo in X e l'altro in Y. In altre parole, i grafi bipartiti sono esattamente quelli che possono essere colorati con due colori (uno per X ed uno per Y).



Un grafo bipartito con un accoppiamento

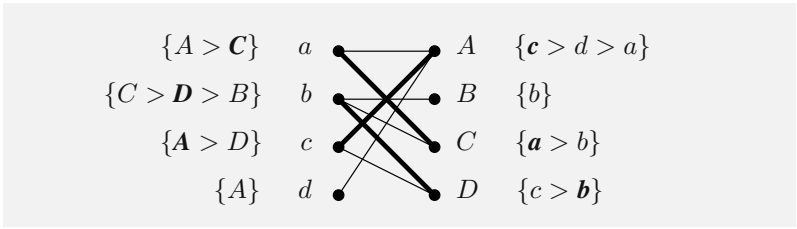
Ora arriviamo ad un concetto importante, gli "accoppiamenti stabili", con un'interpretazione terra-terra. Un *accoppiamento* M in un grafo bipartito $G = (X \cup Y, E)$ è un insieme di spigoli tale che nessuna coppia di spigoli in M ha un estremo comune. Nel grafo mostrato gli spigoli tracciati in grassetto costituiscono un accoppiamento.

Consideriamo un insieme di uomini X ed un insieme di donne Y ed interpretiamo $uv \in E$ come la possibilità che u e v si sposino. Un accoppiamento è allora un matrimonio di massa con nessuna persona in stato di bigamia. Per i nostri scopi ci serve una versione più raffinata (e più realistica?) dell'accoppiamento, suggerita da David Gale e Lloyd S. Shapley. Chiaramente, nella vita reale ogni persona ha preferenze e questo è ciò che aggiungiamo alla configurazione. In $G = (X \cup Y, E)$ supponiamo che per ogni $v \in X \cup Y$ vi sia un ordinamento dell'insieme $N(v)$ dei vertici adiacenti a v, $N(v) = \{z_1 > z_2 > \ldots > z_{d(v)}\}$. Pertanto z_1 è la scelta migliore per v, seguita da z_2, e così via.

Definizione 2. Un accoppiamento M di $G = (X \cup Y, E)$ è detto *stabile* se vale la seguente condizione: ogniqualvolta $uv \in E \backslash M$, $u \in X$, $v \in Y$, allora o $uy \in M$ con $y > v$ in $N(u)$ oppure $xv \in M$ con $x > u$ in $N(v)$, oppure entrambe sono vere.

Nella nostra interpretazione "dalla vita reale" un insieme di matrimoni è stabile se non avviene mai che u e v non siano sposati ma u preferisca v al suo partner (ammesso che ne abbia uno) e v preferisca u al suo compagno (ammesso che ne abbia uno), il che chiaramente creerebbe una situazione instabile.

Prima di dimostrare il nostro secondo risultato diamo un'occhiata al seguente esempio:



Gli spigoli in grassetto costituiscono un accoppiamento stabile. In ogni lista di priorità, la scelta che porta ad un accoppiamento stabile è stampata in grassetto

Si noti che in questo esempio vi è un unico accoppiamento massimo M con quattro spigoli, $M = \{aC, bB, cD, dA\}$, ma M non è stabile (si consideri cA).

Lemma 2. *Un accoppiamento stabile esiste sempre.*

■ **Dimostrazione.** Si consideri il seguente algoritmo. Nel primo stadio tutti gli uomini $u \in X$ si propongono alla loro scelta numero uno. Se una ragazza riceve più di un candidato sceglie quello che preferisce e lo tiene con una cordicella, idem se riceve solo un candidato. Gli uomini rimanenti sono respinti e formano la riserva R. Nel secondo stadio tutti gli uomini in R si propongono alla loro seconda scelta. Le donne paragonano i candidati (incluso quello alla cordicella, se ve ne è uno), scelgono il loro preferito e lo attaccano alla cordicella. Il resto è respinto e forma il nuovo insieme R. Ora gli uomini in R si propongono alla loro scelta seguente, e così via. Un uomo che si è proposto alla sua ultima scelta ed è di nuovo respinto rinuncia ad ulteriori considerazioni (e si toglie dalla riserva). Chiaramente, dopo qualche tempo la riserva R è vuota ed a questo punto l'algoritmo si arresta.

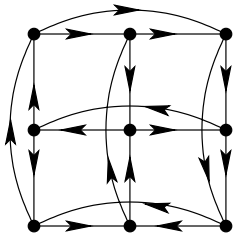
> **Proposizione**. *Quando l'algoritmo si arresta, allora gli uomini tenuti alle cordicelle e le corrispondenti ragazze formano un accoppiamento stabile.*

Si noti innanzitutto che gli uomini alla cordicella di una particolare ragazza vi si trasferiscono in ordine di preferenza crescente (della ragazza) poiché ad ogni passo la ragazza confronta le nuove proposte con il compagno presente e quindi sceglie il nuovo preferito. Pertanto se $uv \in E$ ma $uv \notin M$, allora o u non si è mai proposto a v nel qual caso ha trovato una migliore compagna prima di essere arrivato a v, il che implica che $uy \in M$ con $y > v$ in $N(u)$ oppure u si è proposto a v ma ne è stato respinto, il che implica $xv \in M$ con $x > u$ in $N(v)$. Ma questa è esattamente la condizione di accoppiamento stabile. □

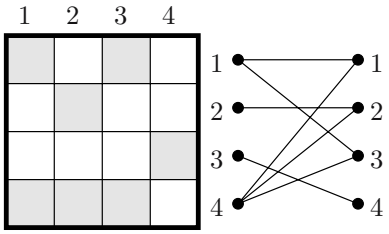
Mettendo insieme i Lemmi 1 e 2 , abbiamo ora la soluzione di Galvin del problema di Dinitz.

Teorema. *Abbiamo $\chi_\ell(S_n) = n$ per ogni n.*

1	2	3
3	1	2
2	3	1



■ **Dimostrazione.** Come prima indichiamo i vertici di S_n con (i, j), $1 \leq i, j \leq n$. Pertanto (i, j) e (r, s) sono adiacenti se e solo se $i = r$ o $j = s$. Si prenda un quadrato latino L con le lettere da $\{1, 2, \ldots, n\}$ e si indichi con



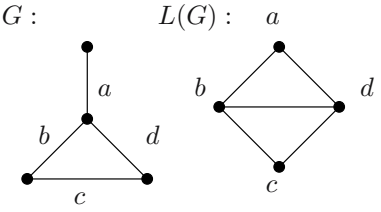
$L(i,j)$ l'elemento nella casella (i,j). Quindi si renda S_n un grafo orientato $\vec{S}_n$ orientando gli spigoli orizzontali $(i,j) \longrightarrow (i,j')$ se $L(i,j) < L(i,j')$ e gli spigoli verticali $(i,j) \longrightarrow (i',j)$ se $L(i,j) > L(i',j)$. Pertanto, orizzontalmente orientiamo dall'elemento più piccolo a quello più grande e verticalmente nell'altro senso. A margine abbiamo un esempio per $n = 3$.

Si noti che otteniamo $d^+(i,j) = n-1$ per ogni (i,j). In effetti, se $L(i,j) = k$, allora $n-k$ caselle nella riga i contengono un elemento maggiore di k, e $k-1$ caselle nella colonna j hanno un elemento minore di k.

In base al Lemma 1 resta da dimostrare che ogni sottografo indotto di $\vec{S}_n$ possiede un nucleo. Si consideri un sottoinsieme $A \subseteq V$ e siano X e Y l'insieme delle righe e quello delle colonne di L. Si associ ad A il grafo bipartito $G = (X \cup Y, A)$, dove ogni $(i,j) \in A$ è rappresentato dallo spigolo ij con $i \in X, j \in Y$. Nell'esempio a margine le caselle di A sono ombreggiate.

L'orientamento su S_n induce naturalmente un ordinamento sugli intorni in $G = (X \cup Y, A)$ ponendo $j' > j$ in $N(i)$ se $(i,j) \longrightarrow (i,j')$ in $\vec{S}_n$ rispettivamente $i' > i$ in $N(j)$ se $(i,j) \longrightarrow (i',j)$. In base al Lemma 2, $G = (X \cup Y, A)$ possiede un accoppiamento stabile M. Questo M, visto come un sottoinsieme di A, è il nucleo desiderato! Per vedere perché, si noti per cominciare che M è indipendente in A in quanto famiglia di spigoli in $G = (X \cup Y, A)$ e questi ultimi non condividono un'estremità i o j. In secondo luogo, se $(i,j) \in A \backslash M$, allora per la definizione di accoppiamento stabile esiste o $(i,j') \in M$ con $j' > j$ oppure $(i',j) \in M$ con $i' > i$, che per $\vec{S}_n$ significa $(i,j) \longrightarrow (i,j') \in M$ oppure $(i,j) \longrightarrow (i',j) \in M$ e la dimostrazione è completa. □

Per terminare la storia spingiamoci un po' più in là. Il lettore può aver notato che il grafo S_n deriva da un grafo bipartito mediante una semplice costruzione. Si prendano il grafo bipartito completo, indicato con $K_{n,n}$, con $|X| = |Y| = n$, e *tutti* gli spigoli tra X e Y. Se consideriamo gli spigoli di $K_{n,n}$ come vertici di un nuovo grafo, congiungendo due di questi vertici se e solo se come spigoli in $K_{n,n}$ essi hanno un estremo in comune, allora otteniamo chiaramente il grafo quadrato S_n. Diciamo che S_n è il *grafo degli archi* di $K_{n,n}$. Ora questa stessa costruzione può essere compiuta su ogni grafo G con il grafo risultante denominato *grafo degli archi* $L(G)$ di G.



Costruzione di un grafo degli archi

In generale, si chiami H un *grafo degli archi* se $H = L(G)$ per un grafo G. Ovviamente, non ogni grafo è un grafo degli archi; un esempio è il grafo $K_{2,4}$ che abbiamo considerato prima, e per il quale abbiamo visto $\chi(K_{2,4}) < \chi_\ell(K_{2,4})$. Ma cosa succede se H *è* un grafo degli archi? Adattando la dimostrazione del nostro teorema si può facilmente dimostrare che $\chi(H) = \chi_\ell(H)$ vale ogniqualvolta H è il grafo degli archi di un grafo *bipartito* ed il metodo può spingersi oltre, verso la verifica della congettura suprema in questo campo:

$$\chi(H) = \chi_\ell(H) \ \textit{vale per ogni grafo degli archi } H\textit{?}$$

Molto poco è noto di questa congettura e le cose sembrano difficili — ma

dopo tutto, anche il problema di Dinitz lo era vent'anni fa.

Bibliografia

[1] P. ERDŐS, A. L. RUBIN & H. TAYLOR: *Choosability in graphs,* Proc. West Coast Conference on Combinatorics, Graph Theory and Computing, Congressus Numerantium **26** (1979), 125-157.

[2] D. GALE & L. S. SHAPLEY: *College admissions and the stability of marriage,* Amer. Math. Monthly **69** (1962), 9-15.

[3] F. GALVIN: *The list chromatic index of a bipartite multigraph,* J. Combinatorial Theory, Ser. B **63** (1995), 153-158.

[4] J. C. M. JANSSEN: *The Dinitz problem solved for rectangles,* Bulletin Amer. Math. Soc. **29** (1993), 243-249.

[5] V. G. VIZING: *Coloring the vertices of a graph in prescribed colours (in Russian),* Metody Diskret. Analiz. **101** (1976), 3-10.

Identità contro biiezioni

Capitolo 29

Si consideri il prodotto infinito $(1+x)(1+x^2)(1+x^3)(1+x^4)\cdots$ e lo si sviluppi in serie $\sum_{n\geq 0} a_n x^n$ nel modo consueto raggruppando i prodotti che danno la stessa potenza x^n. Analizzandoli troviamo per i primi termini

$$\prod_{k\geq 1}(1+x^k) = 1 + x + x^2 + 2x^3 + 2x^4 + 3x^5 + 4x^6 + 5x^7 + \dots . \quad (1)$$

Pertanto abbiamo ad esempio $a_6 = 4$, $a_7 = 5$ e sospettiamo (a ragione) che a_n tenda all'infinito per $n \longrightarrow \infty$.

Guardando al prodotto altrettanto semplice $(1-x)(1-x^2)(1-x^3)(1-x^4)\cdots$ accade qualcosa di inatteso. Sviluppandolo otteniamo

$$\prod_{k\geq 1}(1-x^k) \;=\; 1-x-x^2+x^5+x^7-x^{12}-x^{15}+x^{22}+x^{26}-\dots . \quad (2)$$

Sembra che tutti i coefficienti siano uguali a 1, -1 o 0. Ma è vero questo? E in tal caso, qual è il pattern?

Somme e prodotti infiniti e la loro convergenza hanno giocato un ruolo centrale nell'analisi a partire dall'invenzione del calcolo e contributi all'argomento sono stati forniti da alcuni dei più grandi nomi in questo campo, fra questi Eulero e Ramanujan.

Nello spiegare identità quali (1) e (2), tuttavia, non consideriamo questioni di convergenza — manipoliamo semplicemente i coefficienti. Nel linguaggio del settore, diciamo che affrontiamo serie (e prodotti) di potenze "formali". In questo contesto, mostreremo come argomenti combinatori portino a dimostrazioni eleganti di identità all'apparenza difficili.

La nostra nozione di base è quella di una *partizione* di un numero naturale. Definiamo una *partizione* di n una qualunque somma

$$\lambda : \;\; n \;=\; \lambda_1 + \lambda_2 + \ldots + \lambda_t \quad \text{con} \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t \geq 1.$$

$P(n)$ sarà l'insieme di tutte le partizioni di n, con $p(n) := |P(n)|$, dove poniamo $p(0) = 1$.

Cosa hanno in comune le partizioni con il nostro problema? Ebbene, si consideri il prodotto infinito di serie:

$$(1+x+x^2+x^3+\dots)(1+x^2+x^4+x^6+\dots)(1+x^3+x^6+x^9+\dots)\cdots \quad (3)$$

dove il k-esimo fattore è $(1+x^k+x^{2k}+x^{3k}+\dots)$. Qual è il coefficiente di x^n quando sviluppiamo questo prodotto in una serie $\sum_{n\geq 0} a_n x^n$? Una riflessione dovrebbe convincervi che questo è esattamente il numero di modi

$5 = 5$
$5 = 4 + 1$
$5 = 3 + 2$
$5 = 3 + 1 + 1$
$5 = 2 + 2 + 1$
$5 = 2 + 1 + 1 + 1$
$5 = 1 + 1 + 1 + 1 + 1.$

Le partizioni contate da $p(5) = 7$

di scrivere n come una somma

$$\begin{aligned} n &= n_1 \cdot 1 + n_2 \cdot 2 + n_3 \cdot 3 + \dots \\ &= \underbrace{1 + \dots + 1}_{n_1} + \underbrace{2 + \dots + 2}_{n_2} + \underbrace{3 + \dots + 3}_{n_3} + \dots . \end{aligned}$$

Pertanto il coefficiente non è altro che il numero $p(n)$ delle partizioni di n. Poiché la serie geometrica $1 + x^k + x^{2k} + \dots$ è uguale a $\frac{1}{1-x^k}$, abbiamo dimostrato la nostra prima identità:

$$\prod_{k\geq 1} \frac{1}{1-x^k} = \sum_{n\geq 0} p(n)\, x^n . \tag{4}$$

Inoltre, vediamo dalla nostra analisi che il fattore $\frac{1}{1-x^k}$ tiene conto del contributo di k ad una partizione di n. Pertanto, se tralasciamo $\frac{1}{1-x^k}$ dal prodotto a sinistra di (4), allora k non appare in nessuna partizione a destra. Come esempio otteniamo immediatamente

$$\prod_{i\geq 1} \frac{1}{1-x^{2i-1}} = \sum_{n\geq 0} p_o(n)\, x^n, \tag{5}$$

$6 = 5 + 1$
$6 = 3 + 3$
$6 = 3 + 1 + 1 + 1$
$6 = 1 + 1 + 1 + 1 + 1 + 1$
Le partizioni di 6 in parti dispari: $p_o(6) = 4$

dove $p_o(n)$ è il numero delle partizioni di n di cui tutti gli addendi sono *dispari* e l'affermazione analoga vale quando tutti gli addendi sono *pari*.

Ora dovrebbe essere chiaro quale sarà l'n-esimo coefficiente nel prodotto infinito $\prod_{k\geq 1}(1 + x^k)$. Poiché prendiamo da ogni fattore in (3) o 1 oppure x^k, ciò significa che consideriamo solo le partizioni in cui ogni addendo k appare al massimo una volta. In altre parole, il nostro prodotto originale (1) è sviluppato in

$$\prod_{k\geq 1} (1 + x^k) = \sum_{n\geq 0} p_d(n)\, x^n , \tag{6}$$

$7 = 7$
$7 = 5 + 1 + 1$
$7 = 3 + 3 + 1$
$7 = 3 + 1 + 1 + 1 + 1$
$7 = 1 + 1 + 1 + 1 + 1 + 1 + 1$
$7 = 7$
$7 = 6 + 1$
$7 = 5 + 2$
$7 = 4 + 3$
$7 = 4 + 2 + 1.$
Le partizioni di 7 in coppie dispari risp. distinte : $p_o(7) = p_d(7) = 5$

dove $p_d(n)$ è il numero delle partizioni di n in addendi *distinti*.

Ora il metodo delle serie formali mostra tutta la sua potenza. Poiché $1 - x^2 = (1 - x)(1 + x)$ possiamo scrivere

$$\prod_{k\geq 1} (1 + x^k) = \prod_{k\geq 1} \frac{1 - x^{2k}}{1 - x^k} = \prod_{k\geq 1} \frac{1}{1 - x^{2k-1}}$$

dal momento che tutti i fattori $1 - x^{2i}$ con esponente pari si cancellano. Pertanto, i prodotti infiniti in (5) e (6) sono identici e dunque anche le serie; otteniamo così lo splendido risultato

$$p_o(n) = p_d(n) \qquad \text{per ogni } n \geq 0. \tag{7}$$

Un'uguaglianza di questo tipo richiede una semplice dimostrazione per biiezione — almeno questo è il punto di vista di ogni combinatorialista.

Problema. *Siano $P_o(n)$ e $P_d(n)$ le partizioni di n in addendi dispari e in addendi distinti, rispettivamente: si trovi una biiezione di $P_o(n)$ su $P_d(n)$!*

Diverse biiezioni sono note, ma la seguente dovuta a J. W. L. Glaisher (1907) è forse la più interessante. Sia λ una partizione di n in parti dispari. Raccogliamo gli addendi uguali ed abbiamo

$$\begin{aligned} n &= \underbrace{\lambda_1 + \ldots + \lambda_1}_{n_1} + \underbrace{\lambda_2 + \ldots + \lambda_2}_{n_2} + \ldots + \underbrace{\lambda_t + \ldots + \lambda_t}_{n_t} \\ &= n_1 \cdot \lambda_1 + n_2 \cdot \lambda_2 + \ldots + n_t \cdot \lambda_t. \end{aligned}$$

Ora scriviamo $n_1 = 2^{m_1} + 2^{m_2} + \ldots + 2^{m_r}$ nella sua rappresentazione binaria e facciamo lo stesso per gli altri n_i. La nuova partizione λ' di n è allora

$$\lambda' : \ n = 2^{m_1}\lambda_1 + 2^{m_2}\lambda_1 + \ldots + 2^{m_r}\lambda_1 + 2^{k_1}\lambda_2 + \ldots .$$

Per esempio,
$\lambda : 25 = 5{+}5{+}5{+}3{+}3{+}1{+}1{+}1{+}1$
è trasformata da ϕ in
$\lambda' : 25 = (2{+}1)5 + (2)3 + (4)1$
$= 10 + 5 + 6 + 4$
$= 10 + 6 + 5 + 4$

Dobbiamo controllare che λ' sia in $P_d(n)$, e che $\phi : \lambda \longmapsto \lambda'$ sia effettivamente una biiezione. Entrambe le affermazioni sono semplici da verificare: se $2^a\lambda_i = 2^b\lambda_j$ allora $2^a = 2^b$ poiché λ_i e λ_j sono dispari, e pertanto $\lambda_i = \lambda_j$. Dunque λ' è in $P_d(n)$. Specularmente, quando $n = \mu_1 + \mu_2 + \ldots + \mu_s$ è una partizione in addendi distinti, allora invertiamo la biiezione raccogliendo tutti i μ_i con la stessa potenza massima di 2 e scriviamo le parti dispari con la molteplicità appropriata. A margine è mostrato un esempio.

Scriviamo
$\lambda' : 30 = 12 + 6 + 5 + 4 + 3$
come $30 = 4(3{+}1) + 2(3) + 1(5{+}3)$
$= (1)5 + (4{+}2{+}1)3 + (4)1$
ed otteniamo come $\phi^{-1}(\lambda')$ la partizione
$\lambda : 30 = 5+3+3+3+3+3+3+3+3+1+1+1+1$
in addendi dispari

La manipolazione dei prodotti formali ha pertanto portato all'uguaglianza $p_o(n) = p_d(n)$ per le partizioni, poi verificata attraverso una biiezione. Ora invertiamo questo ragionamento, diamo una dimostrazione della biiezione per le partizioni e deduciamo un'identità. Questa volta il nostro obiettivo è di identificare il pattern nello sviluppo (2).
Si guardi

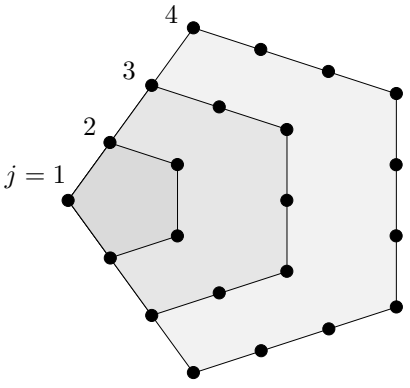
$$1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + x^{22} + x^{26} - x^{35} - x^{40} + \ldots .$$

Gli esponenti (a parte 0) sembrano essere appaiati e prendendo gli esponenti della prima potenza in ogni coppia abbiamo la successione

$$1 \quad 5 \quad 12 \quad 22 \quad 35 \quad 51 \quad 70 \quad \ldots$$

ben nota ad Eulero. Questi sono i *numeri pentagonali* $f(j)$, il cui nome è suggerito dalla figura a margine.

Calcoliamo facilmente $f(j) = \frac{3j^2-j}{2}$ e $\bar{f}(j) = \frac{3j^2+j}{2}$ per l'altro numero in ogni coppia. Riassumendo, congetturiamo, come fece Eulero, che valga la formula seguente.



Numeri pentagonali

Teorema.

$$\prod_{k\geq 1}(1-x^k) \;=\; 1+\sum_{j\geq 1}(-1)^j\Big(x^{\frac{3j^2-j}{2}}+x^{\frac{3j^2+j}{2}}\Big). \qquad (8)$$

Eulero dimostrò questo notevole teorema attraverso calcoli su serie formali, ma noi diamo una prova di biiezione da *Libro*. Innanzitutto, notiamo da (4) che il prodotto $\prod_{k\geq 1}(1-x^k)$ è esattamente l'inverso della nostra serie di partizioni $\sum_{n\geq 0}p(n)x^n$. Pertanto ponendo $\prod_{k\geq 1}(1-x^k) =: \sum_{n\geq 0}c(n)x^n$, troviamo

$$\big(\sum_{n\geq 0}c(n)x^n\big)\cdot\big(\sum_{n\geq 0}p(n)x^n\big) \;=\; 1.$$

Confrontando i coefficienti questo significa che $c(n)$ è l'*unica* successione con $c(0)=1$ e

$$\sum_{k=0}^{n}c(k)p(n-k) \;=\; 0 \qquad \text{per ogni } n\geq 1. \qquad (9)$$

Scrivendo il membro di destra di (8) come $\sum\limits_{j=-\infty}^{\infty}(-1)^j x^{\frac{3j^2+j}{2}}$, dobbiamo dimostrare che

$$c(k) \;=\; \begin{cases} 1 & \text{per } k=\frac{3j^2+j}{2}, \text{quando } j\in\mathbb{Z} \text{ è pari,}\\ -1 & \text{per } k=\frac{3j^2+j}{2}, \text{quando } j\in\mathbb{Z} \text{ è dispari,}\\ 0 & \text{altrimenti}\end{cases}$$

da quest'unica successione. Ponendo $b(j)=\frac{3j^2+j}{2}$ per $j\in\mathbb{Z}$ e sostituendo questi valori in (9), la nostra congettura prende la semplice forma

$$\sum_{j \text{ pari}}p(n-b(j)) \;=\; \sum_{j \text{ dispari}}p(n-b(j)) \qquad \text{per ogni } n,$$

dove ovviamente consideriamo soltanto j con $b(j)\leq n$. Dunque il quadro è impostato: dobbiamo trovare una biiezione

$$\phi:\ \bigcup_{j \text{ pari}}P(n-b(j)) \;\longrightarrow\; \bigcup_{j \text{ dispari}}P(n-b(j)).$$

Anche in questo caso sono state suggerite diverse biiezioni, ma la seguente costruzione di David Bressoud e Doron Zeilberger è di una semplicità stupefacente. Diamo solo la definizione di ϕ (che è, in effetti, un'involuzione) ed invitiamo il lettore a verificare i semplici dettagli.

Per $\lambda : \lambda_1 + \ldots + \lambda_t \in P(n - b(j))$ si ponga

$$\phi(\lambda) \;:=\; \begin{cases} (t+3j-1) + (\lambda_1 - 1) + \ldots + (\lambda_t - 1) & \text{se } t + 3j \geq \lambda_1, \\ (\lambda_2 + 1) + \ldots + (\lambda_t + 1) + \underbrace{1 + \ldots + 1}_{\lambda_1 - t - 3j - 1} & \text{se } t + 3j < \lambda_1, \end{cases}$$

dove tralasciamo i possibili zeri. Si trova che nel primo caso $\phi(\lambda)$ è in $P(n - b(j-1))$ e nel secondo caso in $P(n - b(j+1))$.

Come esempio si consideri $n = 15$, $j = 2$, quindi $b(2) = 7$. La partizione $3 + 2 + 2 + 1$ in $P(15 - b(2)) = P(8)$ è trasformata in $9 + 2 + 1 + 1$, che è in $P(15 - b(1)) = P(13)$

Questo è un ragionamento bellissimo e possiamo dedurne anche di più. Sappiamo già che

$$\prod_{k \geq 1} (1 + x^k) \;=\; \sum_{n \geq 0} p_d(n)\, x^n.$$

Da esperti manipolatori di serie formali notiamo che l'introduzione della nuova variabile y dà

$$\prod_{k \geq 1} (1 + yx^k) \;=\; \sum_{n \geq 0} p_{d,m}(n)\, x^n y^m,$$

dove $p_{d,m}(n)$ conta le partizioni di n in esattamente m addendi distinti. Con $y = -1$ ciò dà

$$\prod_{k \geq 1} (1 - x^k) \;=\; \sum_{n \geq 0} (E_d(n) - O_d(n))\, x^n, \tag{10}$$

dove $E_d(n)$ è il numero di partizioni di n in un numero *pari* di parti distinte, e $O_d(n)$ è il numero di partizioni in un numero *dispari*. Ed ecco la battuta finale. Confrontando (10) con lo sviluppo di Eulero in (8) inferiamo lo splendido risultato

$$E_d(n) - O_d(n) \;=\; \begin{cases} 1 & \text{per } n = \frac{3j^2 \pm j}{2} \text{ quando } j \geq 0 \text{ è pari}, \\ -1 & \text{per } n = \frac{3j^2 \pm j}{2} \text{ quando } j \geq 1 \text{ è dispari}, \\ 0 & \text{altrimenti.} \end{cases}$$

Un esempio per $n = 10$:
$10 = 9 + 1$
$10 = 8 + 2$
$10 = 7 + 3$
$10 = 6 + 4$
$10 = 4 + 3 + 2 + 1$
e
$10 = 10$
$10 = 7 + 2 + 1$
$10 = 6 + 3 + 1$
$10 = 5 + 4 + 1$
$10 = 5 + 3 + 2$,
pertanto $E_d(10) = O_d(10) = 5$.

Questo è, ovviamente, solo l'inizio di una storia più lunga ed ancora in corso. La teoria dei prodotti infiniti è piena di identità inaspettate e con esse le loro controparti biiettive. Gli esempi più celebri sono le cosiddette identità di Rogers-Ramanujan, così chiamate da Leonard Rogers e Srinivasa Ramanujan, in cui il numero 5 gioca un ruolo misterioso:

$$\prod_{k \geq 1} \frac{1}{(1 - x^{5k-4})(1 - x^{5k-1})} \;=\; \sum_{n \geq 0} \frac{x^{n^2}}{(1-x)(1-x^2)\cdots(1-x^n)},$$

$$\prod_{k \geq 1} \frac{1}{(1 - x^{5k-3})(1 - x^{5k-2})} \;=\; \sum_{n \geq 0} \frac{x^{n^2+n}}{(1-x)(1-x^2)\cdots(1-x^n)}.$$

Il lettore è invitato a tradurle nelle seguenti identità di partizioni notate per la prima volta da Percy MacMahon:

Srinivasa Ramanujan

- Sia $f(n)$ il numero delle partizioni di n di cui tutti gli addendi sono della forma $5k+1$ oppure $5k+4$, e $g(n)$ il numero delle partizioni i cui addendi differiscono di almeno 2. Allora $f(n) = g(n)$.

- Sia $r(n)$ il numero di partizioni di n di cui tutti gli addendi sono della forma $5k+2$ o $5k+3$, e sia $s(n)$ il numero di partizioni le cui parti differiscono di almeno 2 e che non contengono 1. Allora $r(n) = s(n)$.

Tutte le dimostrazioni note sulle serie formali delle identità di Rogers-Ramanujan sono piuttosto difficili e per lungo tempo le dimostrazioni di biiezione di $f(n) = g(n)$ e di $r(n) = s(n)$ sono sembrate elusive. Tali dimostrazioni furono infine date nel 1981 da Adriano Garsia e Stephen Milne. Le loro biiezioni sono, tuttavia, molto complicate — non ci sono ancora *Book Proof* in vista.

Bibliografia

[1] G. E. ANDREWS: *The Theory of Partitions,* Encyclopedia of Mathematics and its Applications, Vol. 2, Addison-Wesley, Reading MA 1976.

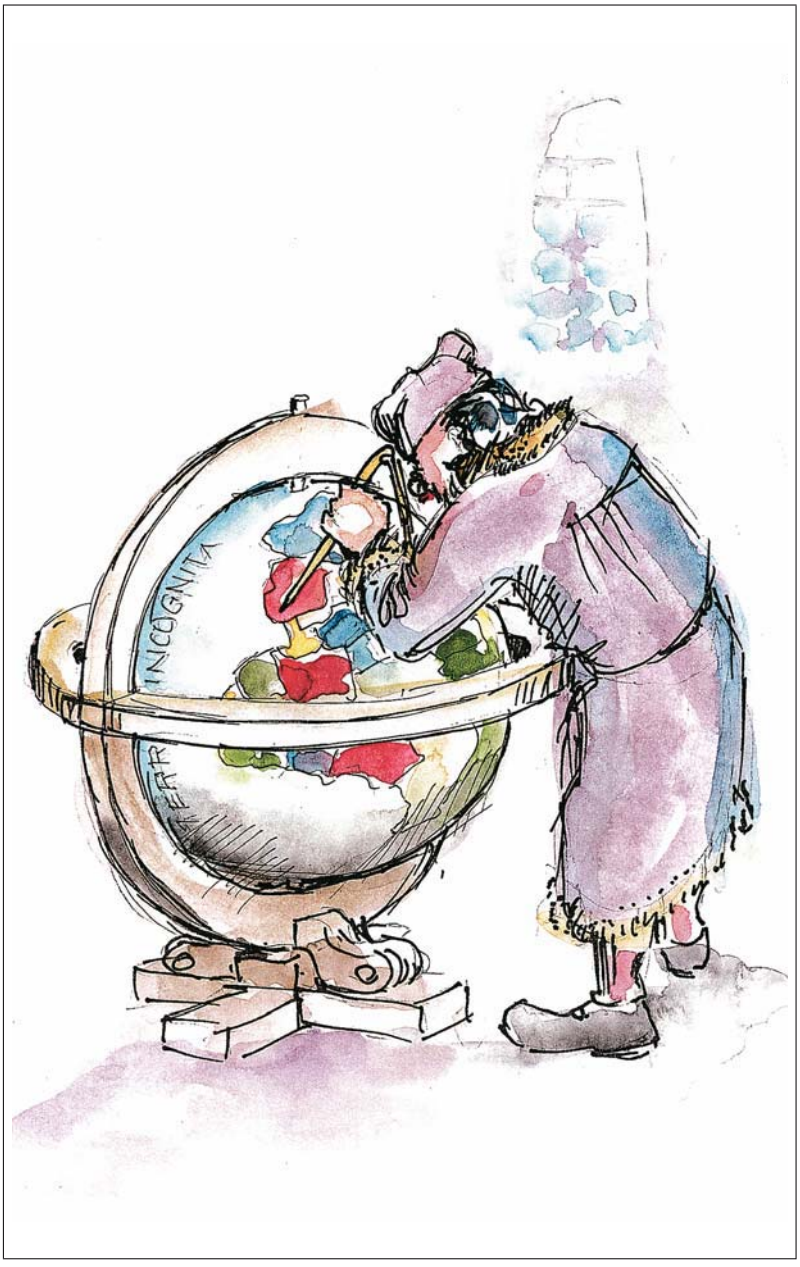
[2] D. BRESSOUD & D. ZEILBERGER: *Bijecting Euler's partitions-recurrence,* Amer. Math. Monthly **92** (1985), 54-55.

[3] A. GARSIA & S. MILNE: *A Rogers-Ramanujan bijection,* J. Combinatorial Theory, Ser. A **31** (1981), 289-339.

[4] S. RAMANUJAN: *Proof of certain identities in combinatory analysis,* Proc. Cambridge Phil. Soc. **19** (1919), 214-216.

[5] L. J. ROGERS: *Second memoir on the expansion of certain infinite products,* Proc. London Math. Soc. **25** (1894), 318-343.

Teoria dei Grafi

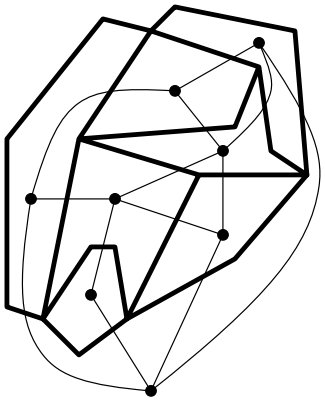


"Il geografo dei quattro colori"

Colorazione di grafi piani con cinque colori

Capitolo 30

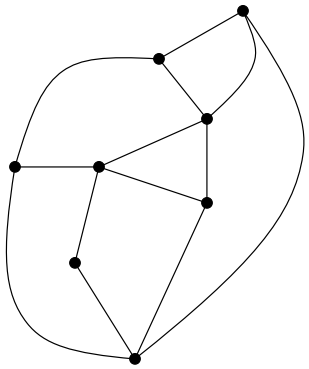
I grafi piani e le loro colorazioni sono stati argomento di ricerche intensive fin dall'inizio della teoria dei grafi per via della loro connessione con il problema dei quattro colori. Come affermato in origine il problema dei quattro colori poneva il quesito se sia sempre possibile colorare le regioni di una carta geografica piana con quattro colori in modo tale che le regioni che condividono un bordo (e non solo un punto) in comune ricevano colori diversi. La figura a fianco mostra che colorare le regioni di una carta è in realtà la stessa operazione che colorare i vertici di un grafo piano. Come nel Capitolo 11 (pagina 71) si ponga un vertice all'interno di ogni regione (inclusa la regione esterna) e si connettano due di tali vertici appartenenti a regioni confinanti mediante un arco attraverso il bordo comune.



Il grafo duale di una carta

Il grafo risultante G, il *grafo duale* della mappa M, è allora un grafo piano e colorare i vertici di G nel senso abituale è equivalente a colorare le regioni di M. Pertanto possiamo concentrarci sulla colorazione di vertici di grafi piani e così faremo d'ora in poi. Si noti che possiamo supporre che G non abbia cicli o archi multipli, poiché questi sono irrilevanti ai fini della colorazione.

Nella lunga e ardua storia delle strategie d'attacco per dimostrare il teorema dei quattro colori molti tentativi si sono avvicinati alla soluzione, ma quello che finalmente riuscì nella dimostrazione di Appel-Haken del 1976 ed anche nella recente dimostrazione di Robertson, Sanders, Seymour e Thomas del 1997 fu una combinazione di idee molto vecchie (risalenti al diciannovesimo secolo) con la potenza di calcolo dei computer dei giorni nostri. Venticinque anni dopo la dimostrazione originale, la situazione resta pressapoco la stessa, nessuna dimostrazione dunque è in vista per il *Libro*.



Questo grafo piano ha 8 vertici, 13 archi e 7 regioni

Dovremmo essere dunque più modesti e ci chiediamo se vi sia una dimostrazione pulita che ogni grafo piano possa essere colorato con cinque colori. Una dimostrazione di questo teorema dei cinque colori fu già fornita da Heawood all'inizio del ventesimo secolo. Lo strumento basilare per la sua dimostrazione (ed in effetti anche per il problema dei quattro colori) fu la formula di Eulero (si veda il Capitolo 11). Chiaramente, quando coloriamo un grafo G possiamo supporre che G sia connesso poiché possiamo colorare le parti connesse separatamente. Un grafo piano divide il piano in un insieme R di regioni (inclusa la regione esterna). La formula di Eulero afferma che per grafi piani connessi $G = (V, E)$ abbiamo sempre

$$|V| - |E| + |R| \;=\; 2.$$

Come riscaldamento, guardiamo come la formula di Eulero possa essere

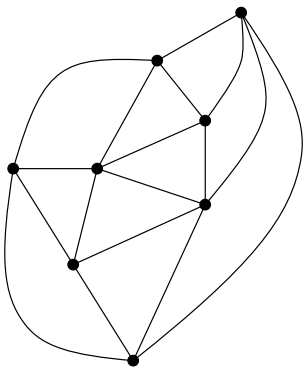
applicata per dimostrare che ogni grafo piano è colorabile con 6 colori. Procediamo per induzione sul numero dei vertici, n. Per valori piccoli di n (in particolare, per $n \leq 6$) questo è ovvio.

Dalla parte (A) della proposizione a pagina 73 sappiamo che G ha un vertice v di grado massimo 5. Si eliminino v e tutti gli archi incidenti in v. Il grafo risultante $G' = G \backslash v$ è un grafo piano su $n - 1$ vertici. Per induzione, esso si può colorare con 6 colori. Poiché v ha al massimo 5 vicini in G, sono usati al massimo 5 colori per tali vicini nella colorazione di G'. Possiamo dunque estendere ogni colorazione con 6 colori di G' ad una colorazione con 6 colori di G assegnando a v un colore che non sia usato per nessuno dei suoi vicini nella colorazione di G'. Pertanto G è effettivamente 6-colorabile.

Esaminiamo ora il numero cromatico della lista dei grafi piani, come discusso nel capitolo sul problema di Dinitz. Chiaramente, il nostro metodo di colorazione con 6 colori funziona anche per le liste di colori (di nuovo non esauriamo mai i colori), pertanto $\chi_\ell(G) \leq 6$ vale per ogni grafo piano G. Erdős, Rubin e Taylor congetturarono nel 1979 che ogni grafo piano abbia come numero cromatico della lista al massimo 5 ed inoltre che vi siano grafi piani G tali che $\chi_\ell(G) > 4$. Avevano ragione su entrambe le cose. Margit Voigt fu la prima a costruire un esempio di grafo piano G tale che $\chi_\ell(G) = 5$ (il suo esempio aveva 238 vertici) ed all'incirca nello stesso tempo Carsten Thomassen diede una dimostrazione veramente stupefacente della congettura sulla colorazione per lista con 5 colori. La sua dimostrazione è un esempio lampante di quello che potete fare quando trovate la giusta ipotesi di induzione. Essa non usa per nulla la formula di Eulero!

Teorema. *Tutti i grafi planari G possono essere colorati per lista con 5 colori:*

$$\chi_\ell(G) \leq 5.$$



Un grafo piano quasi-triangolato

■ **Dimostrazione.** Si noti innanzitutto che aggiungendo archi si può soltanto aumentare il numero cromatico. In altre parole, quando H è un sottografo di G, allora certamente $\chi_\ell(H) \leq \chi_\ell(G)$. Pertanto possiamo supporre che G sia connesso e che tutte le facce limitate di un'immersione abbiamo triangoli come bordi. Chiamiamo un tale grafo *quasi-triangolato*. La validità del teorema per i grafi quasi-triangolati stabilirà l'affermazione per tutti i grafi piani.

Il trucco della dimostrazione è di mostrare la seguente affermazione più forte (che ci permette di usare l'induzione):

> *Sia $G = (V, E)$ un grafo quasi-triangolato e sia B il ciclo che delimita la regione esterna. Formuliamo le seguenti supposizioni sugli insiemi di colori $C(v)$, $v \in V$:*

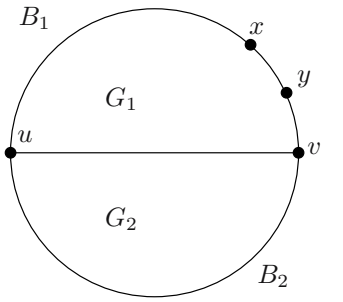
(1) *Due vertici adiacenti x, y di B sono colorati con (diversi) colori α e β.*

(2) $|C(v)| \geq 3$ *per tutti gli altri vertici v di B.*

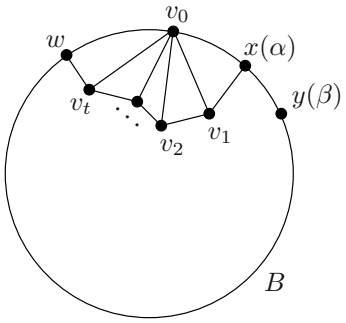
(3) $|C(v)| \geq 5$ *per tutti i vertici v all'interno.*

Allora la colorazione di x, y si può estendere ad una vera e propria colorazione di G scegliendo i colori dalla lista. In particolare, $\chi_\ell(G) \leq 5$.

Per $|V| = 3$ ciò è ovvio, poiché per il solo vertice non colorato v abbiamo $|C(v)| \geq 3$, dunque vi è un colore disponibile. Ora procediamo per induzione.



Caso 1: Supponiamo che B abbia una corda, ovvero, un arco non in B che congiunge due vertici $u, v \in B$. Il sottografo G_1, che è limitato da $B_1 \cup \{uv\}$ e contiene x, y, u e v, è quasi-triangolato e pertanto ha una colorazione per lista con 5 colori, per induzione. Supponiamo che in questa colorazione i vertici u e v ricevano i colori γ e δ. Ora guardiamo alla parte in basso G_2 limitata da B_2 e uv. Considerando u, v come pre-colorata, vediamo che le ipotesi di induzione sono soddisfatte anche per G_2. Pertanto G_2 può essere colorato per lista con 5 colori usando i colori a disposizione, e dunque lo stesso vale per G.



Caso 2: Supponiamo che B non abbia corde. Sia v_0 il vertice dall'altra parte del vertice α-colorato x su B, e siano $x, v_1, \ldots, v_t, w$ i vicini di v_0. Poiché G è quasi-triangolato abbiamo la situazione mostrata nella figura.

Si costruisca il grafo quasi-triangolato $G' = G \backslash v_0$ eliminando da G il vertice v_0 e tutti gli archi che derivano da v_0. Questo G' ha un bordo esterno $B' = (B \backslash v_0) \cup \{v_1, \ldots, v_t\}$. Poiché $|C(v_0)| \geq 3$ per l'ipotesi (2), esistono due colori γ, δ in $C(v_0)$ diversi da α. Ora sostituiamo ogni insieme di colori $C(v_i)$ con $C(v_i) \backslash \{\gamma, \delta\}$, tenendo gli insiemi di colori originali per tutti gli altri vertici in G'. Allora G' soddisfa chiaramente tutte le ipotesi ed è pertanto colorabile per lista con 5 colori, per induzione. Scegliendo γ o δ per v_0, diverse dal colore di w, possiamo estendere la colorazione per lista di G' a tutto G. □

Pertanto, il teorema della colorazione per lista a 5 colori è dimostrato, ma la storia non è ancora finita. Una congettura più forte affermava che il numero cromatico della lista di un grafo piano G vale al massimo uno più del numero cromatico ordinario, ovvero:

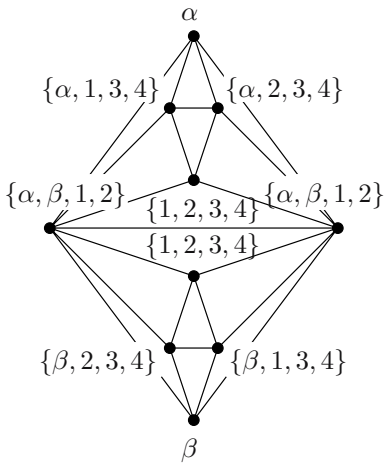
$\chi_\ell(G) \leq \chi(G) + 1$ *per ogni grafo piano G ?*

Poiché $\chi(G) \leq 4$ in base al teorema dei quattro colori, abbiamo tre casi:

Caso I: $\chi(G) = 2 \implies \chi_\ell(G) \leq 3$

Caso II: $\chi(G) = 3 \implies \chi_\ell(G) \leq 4$

Caso III: $\chi(G) = 4 \implies \chi_\ell(G) \leq 5$.



Il risultato di Thomassen sistema il Caso III, mentre il Caso I fu dimostrato con un argomento ingegnoso (e molto più sofisticato) da Alon e Tarsi. Inoltre, vi sono grafi piani G tali che $\chi(G) = 2$ e $\chi_\ell(G) = 3$, ad esempio il grafo $K_{2,4}$ che abbiamo considerato nel capitolo sul problema di Dinitz.

Ma come la mettiamo con il Caso II? Qui la congettura cade, come dimostrato per la prima volta da Margit Voigt per un grafo costruito precedentemente da Shai Gutner. Il suo grafo su 130 vertici si può ottenere come segue. Prima consideriamo il "doppio ottaedro" (si veda la figura), il quale è chiaramente colorabile con 3 colori. Siano $\alpha \in \{5,6,7,8\}$ e $\beta \in \{9,10,11,12\}$ e si considerino le liste date nella figura. Siete invitati a controllare che con queste liste una colorazione non è possibile. Ora prendete 16 copie di questo grafo ed identificate tutti i vertici in alto e quelli in basso. Ciò dà un grafo su $16 \cdot 8 + 2 = 130$ vertici che è ancora piano e colorabile con 3 colori. Assegnamo $\{5,6,7,8\}$ al vertice in alto e $\{9,10,11,12\}$ al vertice in basso, con le liste interne corrispondenti alle 16 coppie (α,β), $\alpha \in \{5,6,7,8\}$, $\beta \in \{9,10,11,12\}$. Per ogni scelta di α e β otteniamo pertanto un sottografo come nella figura e quindi una colorazione per lista del grafo grande non è possibile.

Modificando un altro degli esempi di Gutner, Voigt e Wirth giunsero ad un grafo piano ancora più piccolo con 75 vertici e $\chi = 3$, $\chi_\ell = 5$, che inoltre usa solo il numero minimo di 5 colori nelle liste combinate. Il record attuale è di 63 vertici.

Bibliografia

[1] N. ALON & M. TARSI: *Colorings and orientations of graphs,* Combinatorica **12** (1992), 125-134.

[2] P. ERDŐS, A. L. RUBIN & H. TAYLOR: *Choosability in graphs,* Proc. West Coast Conference on Combinatorics, Graph Theory and Computing, Congressus Numerantium **26** (1979), 125-157.

[3] S. GUTNER: *The complexity of planar graph choosability,* Discrete Math. **159** (1996), 119-130.

[4] N. ROBERTSON, D. P. SANDERS, P. SEYMOUR & R. THOMAS: *The four-colour theorem,* J. Combinatorial Theory, Ser. B **70** (1997), 2-44.

[5] C. THOMASSEN: *Every planar graph is* 5*-choosable,* J. Combinatorial Theory, Ser. B **62** (1994), 180-181.

[6] M. VOIGT: *List colorings of planar graphs,* Discrete Math. **120** (1993), 215-219.

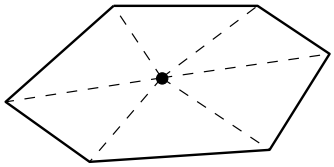
[7] M. VOIGT & B. WIRTH: *On* 3*-colorable non-*4*-choosable planar graphs,* J. Graph Theory **24** (1997), 233-235.

Come sorvegliare un museo

Capitolo 31

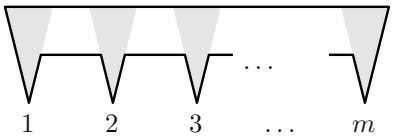
Ecco un problema curioso che fu sollevato da Victor Klee nel 1973. Supponiamo che il direttore di un museo voglia assicurarsi che in ogni momento ogni punto del museo sia controllato da una guardia. Le guardie sono in posti fissi, ma sono in grado di girare. Quante guardie sono necessarie?

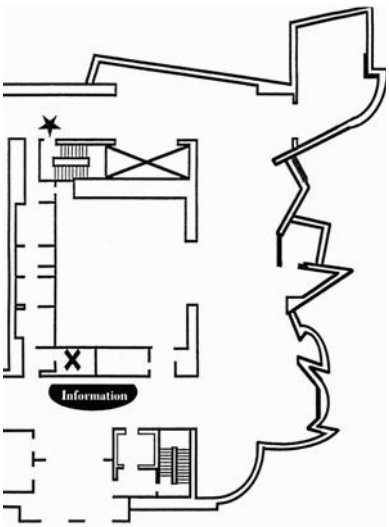
Raffiguriamo i muri di un museo come un poligono composto da n lati. Ovviamente, se il poligono è *convesso*, allora una guardia è sufficiente. In effetti, la guardia può appostarsi in qualunque punto del museo. Tuttavia, in generale, i muri di un museo possono avere la forma di un arbitrario poligono chiuso.



Una sala d'esposizione convessa

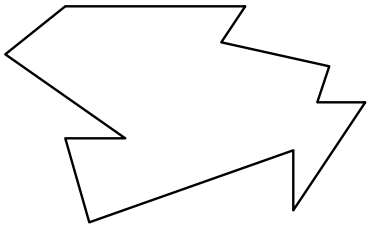
Si consideri un museo a forma di pettine con $n = 3m$ pareti, come raffigurato a lato. È semplice vedere che questo richiede almeno $m = \frac{n}{3}$ guardie. In effetti, ci sono n muri. Ora si noti che il punto 1 può essere osservato solo dalla guardia appostata nel triangolo ombreggiato contenente 1 ed analogamente per gli altri punti $2, 3, \ldots, m$. Poiché tutti questi triangoli sono disgiunti concludiamo che sono necessarie almeno m guardie. Ma m guardie sono anche sufficienti, poiché possono essere poste sulle linee in alto dei triangoli. Tagliando il bordo di una o due pareti, concludiamo che per ogni n vi è un museo con n pareti che richiede $\lfloor \frac{n}{3} \rfloor$ guardie.



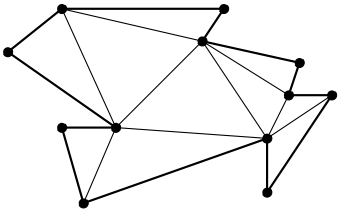


Un vero museo...

Il risultato riportato di seguito afferma che questo è il caso peggiore possibile.



Un museo con $n = 12$ pareti



Una triangolazione del museo

Teorema. *Per ogni museo con n pareti, sono sufficienti $\lfloor \frac{n}{3} \rfloor$ guardie.*

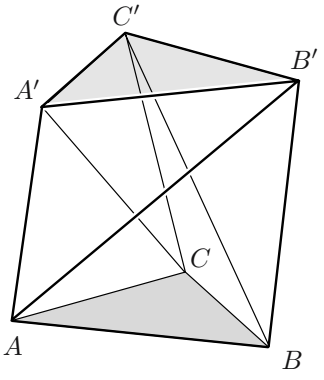
Questo "teorema della galleria d'arte" fu dimostrato per la prima volta da Vašek Chvátal mediante un argomento intelligente; ma ecco una dimostrazione veramente splendida dovuta a Steve Fisk.

■ **Dimostrazione.** Prima di tutto, tracciamo $n - 3$ diagonali che non si incrocino tra gli angoli dei muri al fine di triangolare l'interno. Ad esempio, possiamo tracciare 9 diagonali nel museo raffigurato a margine per produrre una triangolazione. Non importa che triangolazione scegliamo, una qualunque andrà bene. Ora si pensi alla nuova figura come ad un grafo planare con gli angoli come vertici ed i muri e le diagonali come spigoli.

Proposizione. *Questo grafo è colorabile con 3 colori.*

Per $n = 3$ non vi è nulla da dimostrare. Ora per $n > 3$ si scelgano due vertici u e v collegati da una diagonale. Questa diagonale dividerà il grafo in due grafi triangolati più piccoli, entrambi contenenti lo spigolo uv. Per induzione possiamo colorare ogni parte con 3 colori, dove possiamo scegliere il colore 1 per u ed il colore 2 per v in ogni colorazione. Incollando le colorazioni insieme otterremo una colorazione con 3 colori dell'intero grafo.

Il resto è semplice. Dal momento che ci sono n vertici, almeno una delle classi di colori, supponiamo i vertici colorati con 1, contiene al massimo $\lfloor \frac{n}{3} \rfloor$ vertici ed è qui per esempio che poniamo le guardie. Poiché ogni triangolo contiene un vertice di colore 1 deduciamo che ogni triangolo è sorvegliato e pertanto lo è anche l'intero museo. □



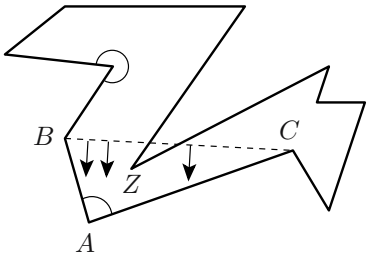
Il poliedro di Schönhardt: gli angoli diedri interni agli spigoli AB', BC' e CA' sono più grandi di 180°

Il lettore astuto potrà aver notato un punto sottile nel nostro ragionamento. Esiste sempre una triangolazione? Probabilmente la prima reazione di ognuno sarà: ovviamente, sì! Ebbene, esiste, ma ciò non è completamente ovvio ed in effetti la generalizzazione naturale a tre dimensioni (partizioni in tetraedri) è falsa! Questo si può notare considerando il *poliedro di Schönhardt*, mostrato a lato. Esso è ottenuto da un prisma triangolare ruotandone il triangolo in alto, cosicché ognuna delle facce quadrilaterali si divide in due triangoli con uno spigolo non convesso. Si provi a triangolare questo poliedro! Noterete che ogni tetraedro che contiene il triangolo in basso deve contenere uno dei tre vertici in alto: ma il tetraedro risultante non sarà contenuto nel poliedro di Schönhardt. Di conseguenza non vi è alcuna triangolazione senza un vertice addizionale.

Per dimostrare che una triangolazione esiste nel caso di un poligono planare non convesso, procediamo per induzione sul numero n dei vertici. Per $n = 3$ il poligono è un triangolo e non vi è nulla da dimostrare. Sia $n \geq 4$.

Per usare l'induzione, tutto quello che dobbiamo produrre è *una* diagonale che dividerà il poligono P in due parti più piccole, in modo tale che una triangolazione del poligono possa essere ottenuta incollando fra loro le triangolazioni delle sue parti.

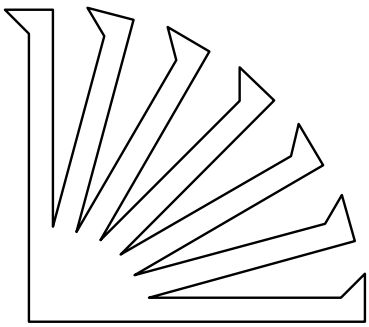
Si definisca un vertice A *convesso* se l'angolo interno al vertice è più piccolo di $180°$. Poiché la somma degli angoli interni di P è $(n-2)180°$, deve esserci un vertice convesso A. In effetti, devono essercene almeno tre: in sostanza questa è un'applicazione del principio del casellario! In alternativa si consideri l'inviluppo convesso del poligono e si noti che tutti i suoi vertici sono convessi anche per il poligono originale.



Ora si considerino i due vertici vicini B e C di A. Se il segmento BC giace interamente in P, allora questa è la nostra diagonale. Altrimenti, il triangolo ABC contiene altri vertici. Si faccia slittare BC verso A fino a toccare l'ultimo vertice Z in ABC. Ora AZ è all'interno di P ed abbiamo una diagonale.

Vi sono molte varianti del teorema del museo. Ad esempio, possiamo voler sorvegliare soltanto le pareti (che sono, dopo tutto, i posti in cui sono appesi i quadri), oppure possiamo appostare tutte le guardie nei vertici. Una variante particolarmente carina (e irrisolta) dice:

> *Si supponga che ogni guardia possa controllare solo una parete del museo, quindi cammini lungo la sua parete e veda tutto quello che si può vedere da qualunque punto lungo tale parete.*
> *Quante "guardie da parete" servono per mantenere il controllo?*



Godfried Toussaint costruì l'esempio di un museo, mostrato qui a lato, per il quale possono essere necessarie $\lfloor \frac{n}{4} \rfloor$ guardie.

Questo poligono ha 28 lati (e, in generale, $4m$ lati) ed il lettore è invitato a controllare che sono necessarie m guardie da parete. Si congettura che, ad eccezione di alcuni valori di n, questo numero sia anche sufficiente, ma manca una dimostrazione, per non parlare di una *Book Proof*.

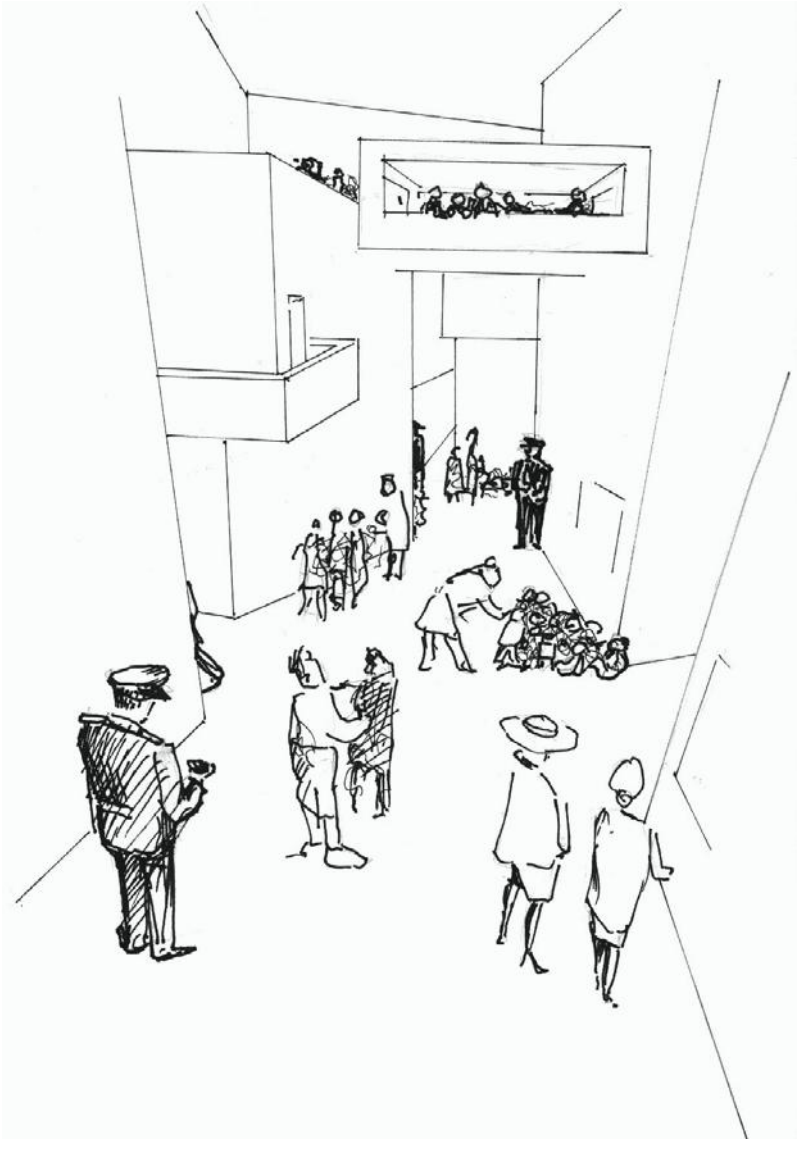
Bibliografia

[1] V. CHVÁTAL: *A combinatorial theorem in plane geometry,* J. Combinatorial Theory, Ser. B **18** (1975), 39-41.

[2] S. FISK: *A short proof of Chvátal's watchman theorem,* J. Combinatorial Theory, Ser. B **24** (1978), 374.

[3] J. O'ROURKE: *Art Gallery Theorems and Algorithms,* Oxford University Press 1987.

[4] E. SCHÖNHARDT: *Über die Zerlegung von Dreieckspolyedern in Tetraeder,* Math. Annalen **98** (1928), 309-312.



"Guardie da museo"
(Un problema di galleria d'arte tridimensionale)

Il teorema dei grafi di Turán

Capitolo 32

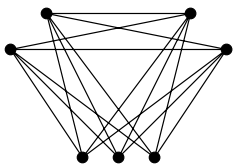
Uno dei risultati fondamentali nella teoria dei grafi è il teorema di Turán del 1941, che diede inizio alla teoria estremale dei grafi. Il teorema di Turán fu riscoperto molte volte con varie dimostrazioni. Ne discuteremo cinque e lasciamo decidere al lettore quale appartenga al *Libro*.

Fissiamo qualche notazione. Consideriamo grafi semplici G sull'insieme dei vertici $V = \{v_1, \ldots, v_n\}$ e degli archi E. Se v_i e v_j sono vicini, allora scriviamo $v_iv_j \in E$. Una *p-clique* in G è un sottografo completo di G su p vertici, indicato da K_p. Paul Turán pose la seguente domanda:

> *Supponiamo che G sia un grafo semplice che non contiene una p-clique. Qual è il più grande numero di archi che G possa avere?*

Paul Turán

Otteniamo prontamente esempi di tali grafi dividendo V in $p-1$ sottoinsiemi disgiunti a due a due $V = V_1 \cup \ldots \cup V_{p-1}$, $|V_i| = n_i$, $n = n_1 + \ldots + n_{p-1}$, unendo due vertici se e solo se giacciono in insiemi distinti V_i, V_j. Indichiamo il grafo risultante con $K_{n_1,\ldots,n_{p-1}}$; esso ha $\sum_{i<j} n_i n_j$ archi. Otteniamo un numero massimo di archi tra questi grafi con un n dato se dividiamo i numeri n_i il più uniformemente possibile, ovvero, se $|n_i - n_j| \leq 1$ per ogni i, j. In effetti, supponiamo che $n_1 \geq n_2 + 2$. Spostando un vertice da V_1 a V_2, otteniamo $K_{n_1-1,n_2+1,\ldots,n_{p-1}}$ che contiene $(n_1-1)(n_2+1) - n_1n_2 = n_1 - n_2 - 1 \geq 1$ più archi che $K_{n_1,n_2,\ldots,n_{p-1}}$. Chiamiamo i grafi $K_{n_1,\ldots,n_{p-1}}$ con $|n_i - n_j| \leq 1$ i *grafi di Turán*. In particolare, se $p-1$ divide n, allora possiamo scegliere $n_i = \frac{n}{p-1}$ per ogni i, ottenendo



Il grafo $K_{2,2,3}$

$$\binom{p-1}{2}\Big(\frac{n}{p-1}\Big)^2 \;=\; \Big(1 - \frac{1}{p-1}\Big)\frac{n^2}{2}$$

archi. Il teorema di Turán afferma ora che questo numero è una maggiorazione per il numero di archi di *ogni* grafo su n vertici senza p-clique.

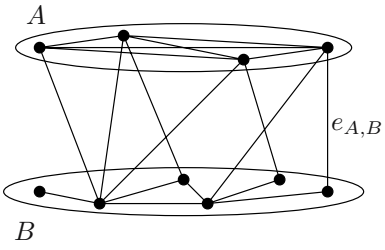
Teorema. *Se un grafo $G = (V, E)$ su n vertici non ha p-clique, $p \geq 2$, allora*

$$|E| \;\leq\; \Big(1 - \frac{1}{p-1}\Big)\frac{n^2}{2}. \tag{1}$$

Per $p = 2$ ciò è banale. Nel primo caso interessante, $p = 3$, il teorema afferma che un grafo senza triangoli su n vertici contiene al massimo $\frac{n^2}{4}$ archi. Dimostrazioni di questo caso speciale erano note prima del risultato

di Turán. Due eleganti dimostrazioni che sfruttano le disuguaglianze sono contenute nel Capitolo 17.

Passiamo ora al caso generale. Le prime due dimostrazioni usano l'induzione e sono dovute rispettivamente a Turán ed a Erdős.

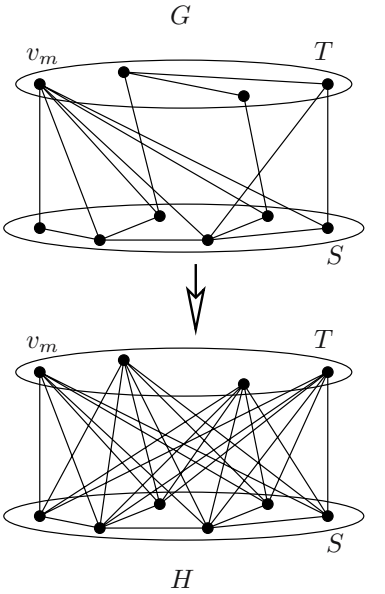


■ **Prima dimostrazione.** Usiamo l'induzione su n. Si può calcolare semplicemente che (1) è vera per $n < p$. Sia G un grafo su $V = \{v_1, \ldots, v_n\}$ senza p-clique con un numero massimo di archi, tale che $n \geq p$. G contiene certamente delle $(p-1)$-clique, poiché altrimenti potremmo aggiungere archi. Sia A una $(p-1)$-clique, e poniamo $B := V \backslash A$.

A contiene $\binom{p-1}{2}$ archi ed ora approssimiamo il numero di archi e_B in B ed il numero di archi $e_{A,B}$ tra A e B. Per induzione, abbiamo $e_B \leq \frac{1}{2}(1-\frac{1}{p-1})(n-p+1)^2$. Poiché G non ha p-clique, ogni $v_j \in B$ è adiacente ad al massimo $p-2$ vertici in A ed otteniamo $e_{A,B} \leq (p-2)(n-p+1)$. Globalmente, ciò dà

$$|E| \leq \binom{p-1}{2} + \frac{1}{2}\Big(1 - \frac{1}{p-1}\Big)(n-p+1)^2 + (p-2)(n-p+1)\,,$$

il che è esattamente $(1-\frac{1}{p-1})\frac{n^2}{2}$. □



■ **Seconda dimostrazione.** Questa dimostrazione fa uso della struttura dei grafi di Turán. Sia $v_m \in V$ un vertice di grado massimo $d_m = \max_{1 \leq j \leq n} d_j$. Si indichi con S l'insieme dei vicini di v_m, $|S| = d_m$ e si ponga $T := V \backslash S$. Poiché G non contiene p-clique e v_m è adiacente a tutti i vertici di S, notiamo che S non contiene $(p-1)$-clique.

Costruiamo ora il seguente grafo H su V (si veda la figura). H corrisponde a G su S e contiene tutti gli archi tra S e T, ma nessuno degli archi all'interno di T. In altre parole, T è un insieme indipendente in H e concludiamo di nuovo che H non ha p-clique. Sia d'_j il grado di v_j in H. Se $v_j \in S$, allora certamente abbiamo $d'_j \geq d_j$ in base alla costruzione di H e per $v_j \in T$ vediamo che $d'_j = |S| = d_m \geq d_j$ in base alla scelta di v_m. Deduciamo che $|E(H)| \geq |E|$ e troviamo che fra tutti i grafi con un numero massimo di archi, ce ne deve essere uno della forma di H. Per induzione, il grafo indotto da S ha al massimo tanti archi quanti quelli di un grafo ben scelto $K_{n_1,\ldots,n_{p-2}}$ su S. Pertanto $|E| \leq |E(H)| \leq E(K_{n_1,\ldots,n_{p-1}})$ con $n_{p-1} = |T|$, il che implica (1). □

Le due dimostrazioni seguenti sono di natura totalmente diversa, usando un argomento di massimizzazione e idee di teoria della probabilità. Esse sono dovute rispettivamente a Motzkin e Straus ed a Alon e Spencer.

■ **Terza dimostrazione.** Si consideri una *distribuzione di probabilità* $\boldsymbol{w} = (w_1, \ldots, w_n)$ sui vertici ovvero un'assegnazione di valori $w_i \geq 0$ ai vertici tale che $\sum_{i=1}^{n} w_i = 1$. Il nostro obiettivo è di massimizzare la funzione

$$f(\boldsymbol{w}) \;=\; \sum_{v_i v_j \in E} w_i w_j.$$

Supponiamo che $\boldsymbol{w}$ sia una distribuzione qualsiasi e siano v_i e v_j una coppia di vertici non adiacenti con pesi positivi w_i, w_j. Sia s_i la somma dei pesi di tutti i vertici adiacenti a v_i e si definisca s_j analogamente per v_j, dove possiamo supporre che $s_i \geq s_j$. Ora spostiamo il peso da v_j a v_i, ovvero, il nuovo peso di v_i è $w_i + w_j$, mentre il peso di v_j scende a 0. Per la nuova distribuzione $\boldsymbol{w}'$ troviamo

$$f(\boldsymbol{w}') \;=\; f(\boldsymbol{w}) + w_j s_i - w_j s_j \;\geq\; f(\boldsymbol{w}).$$

"Spostamento di pesi"

Ripetiamo questo procedimento (riducendo il numero di vertici con un peso positivo di uno ad ogni iterazione) finché non vi siano più vertici non adiacenti di peso positivo. Concludiamo pertanto che vi è una distribuzione ottimale i cui pesi non nulli sono concentrati su una clique, diciamo una k-clique. Ora, se ad esempio $w_1 > w_2 > 0$, si scelga ε con $0 < \varepsilon < w_1 - w_2$ e si trasformino w_1 in $w_1 - \varepsilon$ e w_2 in $w_2 + \varepsilon$. La nuova distribuzione $\boldsymbol{w}'$ soddisfa $f(\boldsymbol{w}') = f(\boldsymbol{w}) + \varepsilon(w_1 - w_2) - \varepsilon^2 > f(\boldsymbol{w})$ ed inferiamo che il valore massimo di $f(\boldsymbol{w})$ è raggiunto per $w_i = \frac{1}{k}$ su una k-clique, altrimenti $w_i = 0$. Poiché una k-clique contiene $\frac{k(k-1)}{2}$ archi, otteniamo

$$f(\boldsymbol{w}) \;=\; \frac{k(k-1)}{2}\frac{1}{k^2} \;=\; \frac{1}{2}\Big(1 - \frac{1}{k}\Big).$$

Poiché questa espressione è crescente rispetto a k, il meglio che si può fare è porre $k = p - 1$ (poiché G non ha p-clique). Dunque concludiamo

$$f(\boldsymbol{w}) \;\leq\; \frac{1}{2}\Big(1 - \frac{1}{p-1}\Big)$$

per *ogni* distribuzione $\boldsymbol{w}$. In particolare, questa disuguaglianza è valida per la distribuzione *uniforme* data da $w_i = \frac{1}{n}$ per ogni i. Pertanto troviamo

$$\frac{|E|}{n^2} \;=\; f\Big(w_i = \frac{1}{n}\Big) \;\leq\; \frac{1}{2}\Big(1 - \frac{1}{p-1}\Big),$$

che è esattamente (1). □

■ **Quarta dimostrazione.** Questa volta usiamo dei concetti tratti dalla teoria delle probabilità. Sia G un grafo arbitrario sull'insieme dei vertici $V = \{v_1, \ldots, v_n\}$. Si indichi il grado di v_i con d_i e si indichi con $\omega(G)$ il numero dei vertici in una clique massima, detto il *numero di clique* di G.

Proposizione. *Abbiamo* $\omega(G) \;\geq\; \displaystyle\sum_{i=1}^{n} \frac{1}{n - d_i}$.

Scegliamo una permutazione arbitraria $\pi = v_1 v_2 \ldots v_n$ dell'insieme dei vertici V, dove ogni permutazione è prevista con la stessa probabilità $\frac{1}{n!}$, e quindi si consideri il seguente insieme C_π. Poniamo v_i in C_π se e solo se v_i è adiacente a tutti i v_j $(j < i)$ che precedono v_i. Per definizione, C_π è una clique in G. Sia $X = |C_\pi|$ la variabile aleatoria corrispondente. Abbiamo

$X = \sum_{i=1}^{n} X_i$, dove X_i è la variabile aleatoria indicante il vertice v_i, ovvero, $X_i = 1$ o $X_i = 0$ a seconda che $v_i \in C_\pi$ oppure $v_i \notin C_\pi$. Si noti che v_i appartiene a C_π rispetto alla permutazione $v_1 v_2 \dots v_n$ se e solo se v_i appare *prima* di tutti gli $n - 1 - d_i$ vertici non adiacenti a v_i, o in altre parole, se v_i è il *primo* tra v_i ed i suoi $n - 1 - d_i$ non-vicini. La probabilità che questo accada è $\frac{1}{n-d_i}$, pertanto $EX_i = \frac{1}{n-d_i}$.
Dunque in base alla linearità del valore atteso (si veda a pagina 94) otteniamo

$$E(|C_\pi|) \;=\; EX \;=\; \sum_{i=1}^{n} EX_i \;=\; \sum_{i=1}^{n} \frac{1}{n-d_i}.$$

Di conseguenza, deve esistere una clique di almeno quella dimensione, e questa era la nostra affermazione. Per dedurre il teorema di Turán dall'affermazione, usiamo la disuguaglianza di Cauchy-Schwarz del Capitolo 17,

$$\Big(\sum_{i=1}^{n} a_i b_i\Big)^2 \;\le\; \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{n=1}^{n} b_i^2\Big).$$

Si pongano $a_i = \sqrt{n-d_i}$, $b_i = \frac{1}{\sqrt{n-d_i}}$, allora $a_i b_i = 1$ e dunque

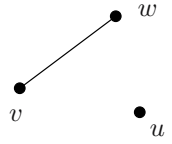
$$n^2 \;\le\; (\sum_{i=1}^{n}(n-d_i))(\sum_{i=1}^{n}\frac{1}{n-d_i}) \;\le\; \omega(G)\sum_{i=1}^{n}(n-d_i). \qquad (2)$$

A questo punto applichiamo l'ipotesi $\omega(G) \le p-1$ del teorema di Turán. Usando anche $\sum_{i=1}^{n} d_i = 2|E|$ dal capitolo sulla conta doppia, la disuguaglianza (2) conduce a

$$n^2 \;\le\; (p-1)(n^2 - 2|E|),$$

e ciò equivale alla disuguaglianza di Turán. □

Siamo ora pronti per l'ultima dimostrazione, forse la più bella di tutte. La sua origine non è chiara; l'abbiamo avuta da Stephan Brandt, il quale ne ebbe notizia in Oberwolfach. Può darsi che derivi dal "folclore" della teoria dei grafi. Essa consente di ottenere in un colpo solo che il grafo di Turán è in realtà l'unico esempio con un numero massimo di archi. Si può notare che entrambe le dimostrazioni 1 e 2 implicano anch'esse questo risultato più forte.



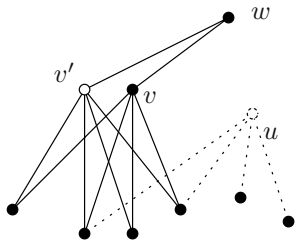
■ **Quinta dimostrazione.** Sia G un grafo su n vertici senza p-clique e con un numero massimo di archi.

> **Proposizione.** *G non contiene tre vertici u, v, w tali che $vw \in E$, ma $uv \notin E$, $uw \notin E$.*

Si supponga il contrario e si considerino i casi seguenti.

Caso 1: $d(u) < d(v)$ o $d(u) < d(w)$.
Possiamo supporre che $d(u) < d(v)$. Allora duplichiamo v ovvero creiamo

un nuovo vertice v' che ha esattamente gli stessi vicini di v (ma vv' non è un arco), eliminiamo u e manteniamo il resto invariato.

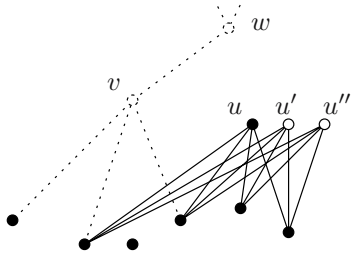


Il nuovo grafo G' è di nuovo privo di p-clique e per il numero degli archi troviamo

$$|E(G')| \;=\; |E(G)| + d(v) - d(u) \;>\; |E(G)|,$$

il che è una contraddizione.

Caso 2: $d(u) \geq d(v)$ e $d(u) \geq d(w)$.
Duplichiamo u due volte ed eliminiamo v e w (come illustrato a margine). Ancora una volta, il nuovo grafo G' non ha p-clique e calcoliamo (il -1 risulta dall'arco vw):



$$|E(G')| \;=\; |E(G)| + 2d(u) - (d(v) + d(w) - 1) \;>\; |E(G)|.$$

Pertanto abbiamo di nuovo una contraddizione.

Un breve ragionamento mostra che l'ipotesi che abbiamo dimostrato è equivalente all'affermazione che

$$u \sim v \;:\Longleftrightarrow\; uv \notin E(G)$$

definisce una relazione di equivalenza. Pertanto G è un grafo completo multipartito, $G = K_{n_1,\ldots,n_{p-1}}$ e così abbiamo terminato. □

Bibliografia

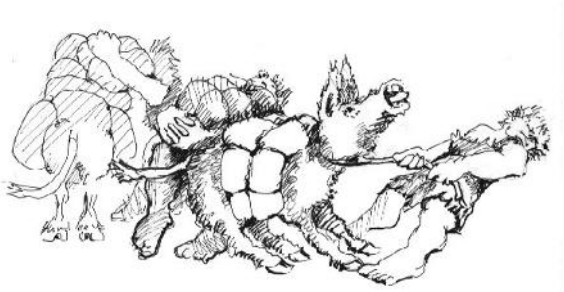
[1] M. AIGNER: *Turán's graph theorem,* Amer. Math. Monthly **102** (1995), 808-816.

[2] N. ALON & J. SPENCER: *The Probabilistic Method,* Wiley Interscience 1992.

[3] P. ERDŐS: *On the graph theorem of Turán (in Hungarian),* Math. Fiz. Lapok **21** (1970), 249-251.

[4] T. S. MOTZKIN & E. G. STRAUS: *Maxima for graphs and a new proof of a theorem of Turán,* Canad. J. Math. **17** (1965), 533-540.

[5] P. TURÁN: *On an extremal problem in graph theory,* Math. Fiz. Lapok **48** (1941), 436-452.



"Pesi più grandi da spostare"

Comunicare senza errori

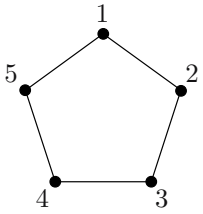
Capitolo 33

Nel 1956, Claude Shannon, il fondatore della teoria dell'informazione, pose una domanda molto interessante:

> *Supponiamo di voler trasmettere dei messaggi attraverso un canale (in cui alcuni simboli possono essere distorti) ad un destinatario. Qual è il tasso di trasmissione massimo affinché il destinatario possa recuperare il messaggio originale senza errori?*

Claude Shannon

Vediamo cosa intendesse Shannon con "canale" e "tasso di trasmissione". Abbiamo un insieme di simboli V ed un messaggio è semplicemente una stringa di simboli da V. Modelliamo il canale come un grafo $G = (V, E)$, dove V è l'insieme dei simboli, ed E è l'insieme dei vertici tra coppie non affidabili di simboli, ovvero, simboli che possono essere confusi durante la trasmissione. Ad esempio, comunicando con il telefono con il linguaggio quotidiano, connettiamo i simboli B e P con un arco dal momento che il destinatario può non essere in grado di distinguerli. Chiamiamo G il *grafo di confusione*.



Il 5-ciclo C_5 giocherà un ruolo prominente nella nostra discussione. In questo esempio, 1 e 2 possono essere confusi, ma non così 1 e 3, etc. Idealmente vorremmo usare tutti e 5 i simboli per la trasmissione, ma dal momento che vogliamo comunicare senza errori possiamo — se inviamo soltanto simboli semplici — usare una sola lettera da ogni coppia che rischia di essere confusa. Pertanto per il 5-ciclo possiamo usare solo due lettere diverse (ogni gruppo di due che non sia collegato da un arco). Nel linguaggio della teoria dell'informazione, ciò significa che per il 5-ciclo raggiungiamo un tasso d'informazione di $\log_2 2 = 1$ (invece del massimo $\log_2 5 \approx 2.32$). È chiaro che in questo modello, per un grafo arbitrario $G = (V, E)$, il meglio che possiamo fare è trasmettere simboli da un insieme indipendente di dimensione massimale. Pertanto il tasso d'informazione, quando si inviano simboli semplici, è $\log_2 \alpha(G)$, dove $\alpha(G)$ è il *numero d'indipendenza* di G.

Vediamo se possiamo aumentare il tasso di informazione usando stringhe più grandi in luogo di simboli semplici. Supponiamo di voler trasmettere stringhe di lunghezza 2. Le stringhe u_1u_2 e v_1v_2 possono essere confuse solo se si verifica una delle tre situazioni seguenti:

- $u_1 = v_1$ e u_2 può essere confuso con v_2,
- $u_2 = v_2$ e u_1 può essere confuso con v_1 o

- $u_1 \neq v_1$ possono essere confusi e $u_2 \neq v_2$ possono essere confusi.

In termini di teoria dei grafi questo si riduce al considerare il *prodotto* $G_1 \times G_2$ di due grafi $G_1 = (V_1, E_1)$ e $G_2 = (V_2, E_2)$. $G_1 \times G_2$ ha come insieme dei vertici $V_1 \times V_2 = \{(u_1, u_2) : u_1 \in V_1, u_2 \in V_2\}$, con $(u_1, u_2) \neq (v_1, v_2)$ connessi da un arco se e solo se $u_i = v_i$ o $u_i v_i \in E_i$ per $i = 1, 2$. Il grafo di confusione per stringhe di lunghezza 2 è pertanto $G^2 = G \times G$, il prodotto del grafo di confusione G per simboli semplici per se stesso. Il tasso di informazione delle stringhe di lunghezza 2 *per simbolo* è dato allora da

$$\frac{\log_2 \alpha(G^2)}{2} \;=\; \log_2 \sqrt{\alpha(G^2)}.$$

Ora, ovviamente, possiamo usare stringhe di ogni lunghezza n. Il grafo di confusione n-esimo, $G^n = G \times G \times \ldots \times G$, ha come insieme dei vertici $V^n = \{(u_1, \ldots, u_n) : u_i \in V\}$ con $(u_1, \ldots, u_n) \neq (v_1, \ldots v_n)$ connesso con un arco se $u_i = v_i$ o $u_i v_i \in E$ per ogni i. Il tasso di informazione per simbolo determinato dalle stringhe di lunghezza n è

$$\frac{\log_2 \alpha(G^n)}{n} \;=\; \log_2 \sqrt[n]{\alpha(G^n)}.$$

Cosa possiamo dire di $\alpha(G^n)$? Ecco una prima osservazione. Sia $U \subseteq V$ un insieme indipendente massimo in G, $|U| = \alpha$. Gli α^n vertici in G^n della forma $(u_1, \ldots, u_n)$, $u_i \in U$ per ogni i, formano chiaramente un insieme indipendente in G^n. Di conseguenza

$$\alpha(G^n) \;\geq\; \alpha(G)^n$$

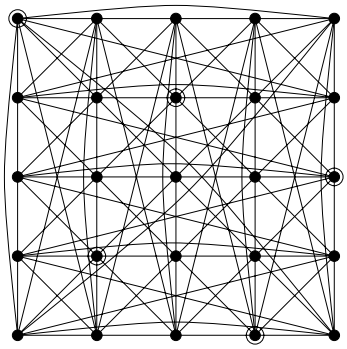
e dunque

$$\sqrt[n]{\alpha(G^n)} \;\geq\; \alpha(G),$$

il che significa che non diminuiamo mai il tasso di informazione usando stringhe più lunghe invece di simboli semplici. Questa, peraltro, è un'idea di base della teoria dei codici: codificando simboli in stringhe più lunghe possiamo rendere la comunicazione senza errori più efficiente. Ignorando il logaritmo arriviamo pertanto alla definizione fondamentale di Shannon: la *capacità senza errori* di un grafo G è data da

$$\Theta(G) \;:=\; \sup_{n \geq 1} \sqrt[n]{\alpha(G^n)},$$

ed il problema di Shannon consisteva nel calcolare $\Theta(G)$, in particolare $\Theta(C_5)$.



Il grafo $C_5 \times C_5$

Esaminiamo C_5. Finora sappiamo che $\alpha(C_5) = 2 \leq \Theta(C_5)$. Considerando il 5-ciclo come rappresentato prima, o il prodotto $C_5 \times C_5$ come disegnato a lato, vediamo che l'insieme $\{(1,1), (2,3), (3,5), (4,2), (5,4)\}$ è indipendente in ${C_5}^2$. Pertanto abbiamo $\alpha({C_5}^2) \geq 5$. Poiché un insieme indipendente può contenere soltanto due vertici provenienti da due righe consecutive, vediamo che $\alpha({C_5}^2) = 5$. Quindi, usando stringhe di lunghezza 2 abbiamo aumentato la minorazione per la capacità a $\Theta(C_5) \geq \sqrt{5}$.

Finora non abbiamo maggiorazioni per la capacità. Per ottenere tali maggiorazioni seguiamo nuovamente le idee originali di Shannon. Innanzitutto abbiamo bisogno della definizione del duale di un insieme indipendente. Ricordiamo che un sottoinsieme $C \subseteq V$ è una *clique* se ogni coppia di vertici di C è collegata da un arco. Pertanto i vertici formano delle clique banali di dimensione 1, gli archi sono le clique di dimensione 2, i triangoli sono clique di dimensione 3, e così via. Sia $\mathcal{C}$ l'insieme delle clique in G. Si consideri una distribuzione di probabilità arbitraria $\boldsymbol{x} = (x_v : v \in V)$ sull'insieme dei vertici, ovvero, $x_v \geq 0$ e $\sum_{v \in V} x_v = 1$. Ad ogni distribuzione $\boldsymbol{x}$ associamo il "valore massimale di una clique"

$$\lambda(\boldsymbol{x}) \;=\; \max_{C \in \mathcal{C}} \sum_{v \in C} x_v,$$

ed infine poniamo

$$\lambda(G) \;=\; \min_{\boldsymbol{x}} \lambda(\boldsymbol{x}) \;=\; \min_{\boldsymbol{x}} \max_{C \in \mathcal{C}} \sum_{v \in C} x_v.$$

Per la precisione dovremmo usare inf invece di min, ma il minimo esiste poiché $\lambda(\boldsymbol{x})$ è continuo sull'insieme compatto di tutte le distribuzioni.

Si consideri ora un insieme indipendente $U \subseteq V$ di dimensione massima $\alpha(G) = \alpha$. Associata a U definiamo la distribuzione $\boldsymbol{x}_U = (x_v : v \in V)$ ponendo $x_v = \frac{1}{\alpha}$ se $v \in U$ e $x_v = 0$ altrimenti. Poiché ogni clique contiene al massimo un vertice di U, inferiamo che $\lambda(\boldsymbol{x}_U) = \frac{1}{\alpha}$, e pertanto per definizione di $\lambda(G)$

$$\lambda(G) \leq \frac{1}{\alpha(G)} \qquad \text{o} \qquad \alpha(G) \leq \lambda(G)^{-1}.$$

Ciò che Shannon osservò è che $\lambda(G)^{-1}$ è in realtà una maggiorazione per ogni $\sqrt[n]{\alpha(G^n)}$ e pertanto anche per $\Theta(G)$. Per dimostrarlo basta verificare che per i grafi G, H abbiamo

$$\lambda(G \times H) \;=\; \lambda(G)\lambda(H), \tag{1}$$

dal momento che ciò implicherà $\lambda(G^n) = \lambda(G)^n$ e dunque

$$\begin{aligned} \alpha(G^n) &\leq \lambda(G^n)^{-1} = \lambda(G)^{-n} \\ \sqrt[n]{\alpha(G^n)} &\leq \lambda(G)^{-1}. \end{aligned}$$

Per dimostrare (1) usiamo il teorema della dualità della programmazione lineare (si veda [1]) ed otteniamo

$$\lambda(G) \;=\; \min_{\boldsymbol{x}} \max_{C \in \mathcal{C}} \sum_{v \in C} x_v \;=\; \max_{\boldsymbol{y}} \min_{v \in V} \sum_{C \ni v} y_C, \tag{2}$$

dove il membro destro si estende a tutte le distribuzioni di probabilità $\boldsymbol{y} = (y_C : C \in \mathcal{C})$ su $\mathcal{C}$.

Si consideri $G \times H$ e siano $\boldsymbol{x}$ e $\boldsymbol{x}'$ delle distribuzioni che raggiungono i minimi, $\lambda(\boldsymbol{x}) = \lambda(G)$, $\lambda(\boldsymbol{x}') = \lambda(H)$. Nell'insieme dei vertici di $G \times H$

assegnamo il valore $z_{(u,v)} = x_u x'_v$ al vertice (u, v). Poiché $\sum_{(u,v)} z_{(u,v)} = \sum_u x_u \sum_v x'_v = 1$, otteniamo una distribuzione. In seguito osserviamo che le clique massime in $G \times H$ sono della forma $C \times D = \{(u, v) : u \in C, v \in D\}$ dove C e D sono clique in G e H, rispettivamente. Pertanto otteniamo

$$\begin{aligned} \lambda(G \times H) \;\leq\; \lambda(\boldsymbol{z}) \;&=\; \max_{C \times D} \sum_{(u,v) \in C \times D} z_{(u,v)} \\ &=\; \max_{C \times D} \sum_{u \in C} x_u \sum_{v \in D} x'_v \;=\; \lambda(G)\lambda(H) \end{aligned}$$

in base alla definizione di $\lambda(G \times H)$. Nello stesso modo la disuguaglianza inversa $\lambda(G \times H) \geq \lambda(G)\lambda(H)$ è dimostrata usando l'espressione duale per $\lambda(G)$ in (2). Riassumendo possiamo affermare:

$$\Theta(G) \;\leq\; \lambda(G)^{-1},$$

per ogni grafo G.

Applichiamo quanto trovato al 5-ciclo C_5 e, più in generale, all'm-ciclo C_m. Usando la distribuzione uniforme $(\frac{1}{m}, \ldots, \frac{1}{m})$ sui vertici, otteniamo $\lambda(C_m) \leq \frac{2}{m}$, poiché ogni clique contiene al massimo due vertici. Analogamente, scegliendo $\frac{1}{m}$ per gli archi e 0 per i vertici, abbiamo $\lambda(C_m) \geq \frac{2}{m}$ in base all'espressione duale in (2). Concludiamo che $\lambda(C_m) = \frac{2}{m}$ e pertanto

$$\Theta(C_m) \;\leq\; \frac{m}{2}$$

per ogni m. Ora, se m è pari, allora chiaramente $\alpha(C_m) = \frac{m}{2}$ e pertanto anche $\Theta(C_m) = \frac{m}{2}$. Per m dispari, tuttavia, abbiamo $\alpha(C_m) = \frac{m-1}{2}$. Per $m = 3$, C_3 è una clique, e così anche ogni prodotto C_3^n, implicando $\alpha(C_3) = \Theta(C_3) = 1$. Dunque, il primo caso interessante è il 5-ciclo, per il quale sino ad ora sappiamo che

$$\sqrt{5} \;\leq\; \Theta(C_5) \;\leq\; \frac{5}{2}. \tag{3}$$

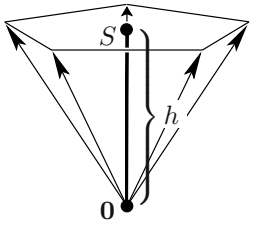
Usando il suo approccio basato sulla programmazione lineare (ed alcune altre idee) Shannon fu in grado di calcolare la capacità di molti grafi e, in particolare, di tutti i grafi con cinque o meno vertici – con la sola eccezione di C_5, dove non potè andare al di là delle maggiorazioni fornite in (3). Le cose rimasero invariate per più di 20 anni finché László Lovász dimostrò con un argomento di una semplicità sbalorditiva che in effetti $\Theta(C_5) = \sqrt{5}$. Un problema combinatorio all'apparenza molto difficile veniva così risolto con una soluzione inattesa ed elegante.

L'idea più innovativa di Lovász fu di rappresentare i vertici v del grafo con dei vettori reali di lunghezza 1 in modo tale che i membri di ogni coppia di vettori appartenenti a vertici non adiacenti in G siano ortogonali. Chiamiamo un tale insieme di vettori una *rappresentazione ortonormale* di G. Chiaramente, tale rappresentazione esiste sempre: si prendano semplicemente i vettori unitari $(1, 0, \ldots, 0)^T$, $(0, 1, 0, \ldots, 0)^T$, …, $(0, 0, \ldots, 1)^T$ di dimensione $m = |V|$.

Per il grafo C_5 possiamo ottenere una rappresentazione ortonormale in $\mathbb{R}^3$ considerando un "ombrello" con cinque stecche $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_5$ di lunghezza unitaria. Ora apriamo l'ombrello (con la punta nell'origine) fino al punto in cui gli angoli tra stecche alternate siano di $90°$.

Lovász procedette poi dimostrando che l'altezza h dell'ombrello, ovvero la distanza tra $\mathbf{0}$ ed S, fornisce la maggiorazione

$$\Theta(C_5) \;\leq\; \frac{1}{h^2}. \tag{4}$$



L'ombrello di Lovász

Un semplice calcolo dà $h^2 = \frac{1}{\sqrt{5}}$; si veda il riquadro a pagina seguente. Da questo segue $\Theta(C_5) \leq \sqrt{5}$ e pertanto $\Theta(C_5) = \sqrt{5}$.

Guardiamo come Lovász procedette per dimostrare la disuguaglianza (4). I suoi risultati erano, in realtà, molto più generali. Si consideri il consueto prodotto interno

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = x_1 y_1 + \ldots + x_s y_s$$

di due vettori $\boldsymbol{x} = (x_1, \ldots, x_s)$, $\boldsymbol{y} = (y_1, \ldots, y_s)$ in $\mathbb{R}^s$. Allora $|\boldsymbol{x}|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = x_1^2 + \ldots + x_s^2$ è il quadrato delle lunghezze $|\boldsymbol{x}|$ di $\boldsymbol{x}$, e l'angolo γ tra $\boldsymbol{x}$ e $\boldsymbol{y}$ è dato da

$$\cos\gamma \;=\; \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{|\boldsymbol{x}||\boldsymbol{y}|}.$$

Pertanto $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ se e solo se $\boldsymbol{x}$ e $\boldsymbol{y}$ sono ortogonali.

I pentagoni e la sezione aurea

Era tradizione che un rettangolo fosse considerato esteticamente gradevole se, dopo aver tagliato un quadrato di lunghezza a, il rettangolo rimanente avesse la stessa forma dell'originale. Le lunghezze dei lati a, b di tale rettangolo dovevano soddisfare $\frac{b}{a} = \frac{a}{b-a}$. Ponendo $\tau := \frac{b}{a}$ per la proporzione, otteniamo $\tau = \frac{1}{\tau-1}$ ovvero $\tau^2 - \tau - 1 = 0$. Risolvendo l'equazione quadratica otteniamo la *sezione aurea* $\tau = \frac{1+\sqrt{5}}{2} \approx 1.6180$.

Si consideri ora un pentagono regolare di lato a e sia d la lunghezza delle sue diagonali. Era già noto ad Euclide (Libro XIII,8) che $\frac{d}{a} = \tau$, e che il punto di intersezione di due diagonali divide le diagonali nella sezione aurea.

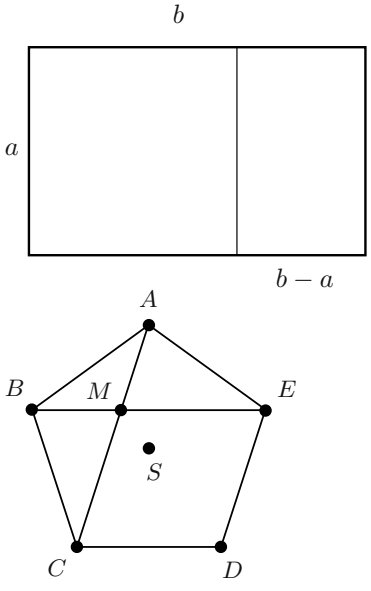
Ecco la *Book Proof* di Euclide. Poiché la somma totale degli angoli del pentagono è 3π, l'angolo ad ogni vertice è uguale a $\frac{3\pi}{5}$. Ne segue che $\sphericalangle ABE = \frac{\pi}{5}$, poiché ABE è un triangolo isoscele. Ciò, a sua volta, implica $\sphericalangle AMB = \frac{3\pi}{5}$, e concludiamo che i triangoli ABC e AMB sono simili. Il quadrilatero $CMED$ è un rombo poiché i lati opposti sono paralleli (si guardino gli angoli) e pertanto $|MC| = a$ e dunque $|AM| = d - a$. In base alla similitudine di ABC ed AMB concludiamo

$$\frac{d}{a} = \frac{|AC|}{|AB|} = \frac{|AB|}{|AM|} = \frac{a}{d-a} = \frac{|MC|}{|MA|} = \tau.$$

Ma c'è dell'altro in arrivo. Per la distanza s di un vertice dal centro del pentagono S, il lettore è invitato a dimostrare la relazione $s^2 = \frac{d^2}{\tau+2}$ (si noti che BS taglia la diagonale AC ad un angolo retto e la dimezza).

Per finire la nostra escursione nella geometria, si consideri ora l'ombrello con il pentagono regolare in cima. Poiché le stecche alterne (di lunghezza 1) formano un angolo retto, il teorema di Pitagora ci dà $d = \sqrt{2}$, e pertanto $s^2 = \frac{2}{\tau+2} = \frac{4}{\sqrt{5}+5}$. Pertanto, ancora usando Pitagora, troviamo per l'altezza $h = |OS|$ il risultato promesso

$$h^2 \;=\; 1 - s^2 \;=\; \frac{1+\sqrt{5}}{\sqrt{5}+5} \;=\; \frac{1}{\sqrt{5}}.$$



Ora ci dirigiamo verso una maggiorazione "$\Theta(G) \leq \sigma_T^{-1}$" per la capacità di Shannon di un grafo qualsiasi G che abbia una rappresentazione ortonormale particolarmente "simpatica". Per far questo sia $T = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ una rappresentazione ortonormale di G in $\mathbb{R}^s$, in cui $\boldsymbol{v}^{(i)}$ corrisponde al vertice v_i. Supponiamo inoltre che tutti i vettori $\boldsymbol{v}^{(i)}$ formino lo *stesso* angolo ($\neq 90°$) con il vettore $\boldsymbol{u} := \frac{1}{m}(\boldsymbol{v}^{(1)} + \ldots + \boldsymbol{v}^{(m)})$. In modo equivalente, supponiamo che il prodotto interno

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle \;=\; \sigma_T$$

abbia lo stesso valore $\sigma_{_T} \neq 0$ per ogni i. Chiamiamo questo valore $\sigma_{_T}$ la *costante* della rappresentazione T. Per l'ombrello di Lovász che rappresenta C_5 la condizione $\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle = \sigma_{_T}$ è certamente valida, essendo $\boldsymbol{u} = \vec{OS}$.

Ora procediamo nei tre passi seguenti.

(A) Si consideri una distribuzione di probabilità $\boldsymbol{x} = (x_1, \ldots, x_m)$ su V e si ponga

$$\mu(\boldsymbol{x}) \;:=\; |x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}|^2$$

e

$$\mu_{_T}(G) \;:=\; \inf_{\boldsymbol{x}} \mu(\boldsymbol{x}).$$

Sia U un insieme indipendente massimo in G tale che $|U| = \alpha$ e si definisca $\boldsymbol{x}_U = (x_1, \ldots, x_m)$ con $x_i = \frac{1}{\alpha}$ se $v_i \in U$ e $x_i = 0$ altrimenti. Poiché tutti i vettori $\boldsymbol{v}^{(i)}$ hanno lunghezza unitaria e $\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(j)} \rangle = 0$ per ogni coppia di vertici non adiacenti, deduciamo

$$\mu_{_T}(G) \;\leq\; \mu(\boldsymbol{x}_U) \;=\; \Big| \sum_{i=1}^{m} x_i \boldsymbol{v}^{(i)} \Big|^2 \;=\; \sum_{i=1}^{m} x_i^2 = \alpha \frac{1}{\alpha^2} = \frac{1}{\alpha}.$$

Pertanto abbiamo $\mu_{_T}(G) \leq \alpha^{-1}$ e dunque

$$\alpha(G) \;\leq\; \frac{1}{\mu_{_T}(G)}.$$

(B) Calcoliamo ora $\mu_{_T}(G)$. Abbiamo bisogno della disuguaglianza di Cauchy-Schwarz

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle^2 \;\leq\; |\boldsymbol{a}|^2 \, |\boldsymbol{b}|^2$$

per vettori $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^s$. Applicata a $\boldsymbol{a} = x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}$ e $\boldsymbol{b} = \boldsymbol{u}$, la disuguaglianza dà

$$\langle x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}, \boldsymbol{u} \rangle^2 \;\leq\; \mu(\boldsymbol{x}) \, |\boldsymbol{u}|^2. \tag{5}$$

In base alla nostra supposizione che $\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle = \sigma_{_T}$ per ogni i, abbiamo

$$\langle x_1 \boldsymbol{v}^{(1)} + \ldots + x_m \boldsymbol{v}^{(m)}, \boldsymbol{u} \rangle \;=\; (x_1 + \ldots + x_m) \, \sigma_{_T} \;=\; \sigma_{_T}$$

per *ogni* distribuzione $\boldsymbol{x}$. Pertanto, in particolare, questo deve valere per la distribuzione uniforme $(\frac{1}{m}, \ldots, \frac{1}{m})$, il che implica $|\boldsymbol{u}|^2 = \sigma_{_T}$. Dunque (5) si riduce a

$$\sigma_{_T}^2 \;\leq\; \mu(\boldsymbol{x}) \, \sigma_{_T} \qquad \text{o} \qquad \mu_{_T}(G) \;\geq\; \sigma_{_T}.$$

D'altro canto, per $\boldsymbol{x} = (\frac{1}{m}, \ldots, \frac{1}{m})$ otteniamo

$$\mu_{_T}(G) \;\leq\; \mu(\boldsymbol{x}) \;=\; |\tfrac{1}{m}(\boldsymbol{v}^{(1)} + \ldots + \boldsymbol{v}^{(m)})|^2 \;=\; |\boldsymbol{u}|^2 \;=\; \sigma_{_T},$$

dunque abbiamo dimostrato

$$\mu_{_T}(G) \;=\; \sigma_{_T}. \tag{6}$$

Riassumendo, abbiamo stabilito la disuguaglianza

$$\alpha(G) \;\leq\; \frac{1}{\sigma_{_T}} \tag{7}$$

per *ogni* rappresentazione ortonormale T con costante $\sigma_{_T}$.

(C) Per estendere questa disuguaglianza a $\Theta(G)$, procediamo come prima. Si consideri di nuovo il prodotto $G \times H$ di due grafi. Abbiamo G e H come rappresentazioni ortonormali R ed S in $\mathbb{R}^r$ e $\mathbb{R}^s$, rispettivamente, con costanti $\sigma_{_R}$ e $\sigma_{_S}$. Sia $\boldsymbol{v} = (v_1, \ldots, v_r)$ un vettore in R e $\boldsymbol{w} = (w_1, \ldots, w_s)$ un vettore in S. Al vertice in $G \times H$ corrispondente alla coppia $(\boldsymbol{v}, \boldsymbol{w})$ associamo il vettore

$$\boldsymbol{v}\boldsymbol{w}^T := (v_1w_1, \ldots, v_1w_s, v_2w_1, \ldots, v_2w_s, \ldots, v_rw_1, \ldots, v_rw_s) \in \mathbb{R}^{rs}.$$

Si verifica subito che $R \times S := \{\boldsymbol{v}\boldsymbol{w}^T : \boldsymbol{v} \in R, \boldsymbol{w} \in S\}$ è una rappresentazione ortonormale di $G \times H$ con costante $\sigma_{_R}\sigma_{_S}$. Pertanto in base a (6) otteniamo

$$\mu_{_{R\times S}}(G \times H) \;\;=\;\; \mu_{_R}(G)\mu_{_S}(H).$$

Per $G^n = G \times \ldots \times G$ e la rappresentazione T con costante $\sigma_{_T}$ ciò significa

$$\mu_{_{T^n}}(G^n) \;=\; \mu_{_T}(G)^n \;=\; \sigma_{_T}^n$$

ed in base a (7) otteniamo

$$\alpha(G^n) \;\leq\; \sigma_{_T}^{-n}, \qquad \sqrt[n]{\alpha(G^n)} \;\leq\; \sigma_{_T}^{-1}.$$

"Ombrelli con cinque stecche"

Mettendo tutto insieme abbiamo dunque completato il ragionamento di Lovász:

Teorema. *Ogniqualvolta* $T = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ *sia una rappresentazione ortonormale di* G *con costante* $\sigma_{_T}$*, allora*

$$\Theta(G) \;\leq\; \frac{1}{\sigma_{_T}}. \tag{8}$$

Guardando l'ombrello di Lovász, abbiamo $\boldsymbol{u} = (0, 0, h{=}\frac{1}{\sqrt[4]{5}})^T$ e pertanto $\sigma = \langle \boldsymbol{v}^{(i)}, \boldsymbol{u}\rangle = h^2 = \frac{1}{\sqrt{5}}$, il che dà $\Theta(C_5) \leq \sqrt{5}$. Dunque il problema di Shannon è risolto.

Portiamo la nostra discussione un po' oltre. Vediamo da (8) che più grande è $\sigma_{_T}$ per una rappresentazione di G, migliore è la maggiorazione che avremo per $\Theta(G)$. Ecco un metodo che ci dà una rappresentazione ortonormale per *ogni* grafo G. A $G = (V, E)$ associamo la *matrice di adiacenza* $A = (a_{ij})$, definita nel modo seguente: sia $V = \{v_1, \ldots, v_m\}$; allora poniamo

$$a_{ij} \;:=\; \begin{cases} 1 & \text{se } v_iv_j \in E \\ 0 & \text{altrimenti.} \end{cases}$$

A è una matrice simmetrica reale con 0 sulla diagonale principale.

Ci servono quindi due proprietà di algebra lineare. Prima di tutto, essendo simmetrica, A ha m autovalori reali $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ (alcuni dei quali possono essere uguali) e la somma degli autovalori è uguale alla somma degli elementi diagonali di A, ovvero 0. Pertanto l'autovalore più piccolo deve essere negativo (eccetto nel caso banale in cui G non abbia archi). Sia $p = |\lambda_m| = -\lambda_m$ il valore assoluto dell'autovalore più piccolo e si consideri la matrice

$$M \;:=\; I \,+\, \frac{1}{p}\,A,$$

dove I indica la matrice identità $(m \times m)$. M ha come autovalori $1 + \frac{\lambda_1}{p} \geq 1 + \frac{\lambda_2}{p} \geq \ldots \geq 1 + \frac{\lambda_m}{p} = 0$. Ora citiamo il secondo risultato (il teorema degli assi principali dell'algebra lineare): se $M = (m_{ij})$ è una matrice simmetrica reale con tutti gli autovalori ≥ 0, allora ci sono dei vettori $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)} \in \mathbb{R}^s$ per $s = \text{rank}(M)$, tali che

$$m_{ij} \;=\; \langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(j)} \rangle \qquad (1 \leq i, j \leq m).$$

In particolare, per $M = I + \frac{1}{p}A$ otteniamo

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(i)} \rangle \;=\; m_{ii} = 1 \qquad \text{per ogni } i$$

e

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(j)} \rangle \;=\; \frac{1}{p} a_{ij} \qquad \text{per } i \neq j.$$

Poiché $a_{ij} = 0$ ogniqualvolta $v_i v_j \notin E$, vediamo che i vettori $\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}$ formano effettivamente una rappresentazione ortonormale di G.

Applichiamo, infine, questa costruzione agli m-cicli C_m per $m \geq 5$ dispari. Qui si calcola facilmente $p = |\lambda_{\min}| = 2\cos\frac{\pi}{m}$ (si veda il riquadro nella pagina seguente). Ogni riga della matrice di adiacenza contiene due 1, il che implica che la somma di ogni riga della matrice M dà $1 + \frac{2}{p}$. Per la rappresentazione $\{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ questo significa

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{v}^{(1)} + \ldots + \boldsymbol{v}^{(m)} \rangle \;=\; 1 + \frac{2}{p} \;=\; 1 + \frac{1}{\cos\frac{\pi}{m}}$$

e quindi

$$\langle \boldsymbol{v}^{(i)}, \boldsymbol{u} \rangle \;=\; \frac{1}{m}(1 + (\cos\tfrac{\pi}{m})^{-1}) \;=\; \sigma$$

per ogni i. Possiamo dunque applicare il nostro risultato principale (8) e concludere

$$\Theta(C_m) \;\leq\; \frac{m}{1 + (\cos\frac{\pi}{m})^{-1}} \qquad (\text{per } m \geq 5 \text{ dispari}). \tag{9}$$

Si noti che, poiché $\cos\frac{\pi}{m} < 1$, la maggiorazione (9) è migliore di quella $\Theta(C_m) \leq \frac{m}{2}$ che abbiamo trovato prima. Si noti inoltre che $\cos\frac{\pi}{5} = \frac{\tau}{2}$, dove $\tau = \frac{\sqrt{5}+1}{2}$ è la sezione aurea. Pertanto per $m = 5$ otteniamo ancora

$$\Theta(C_5) \leq \frac{5}{1 + \frac{4}{\sqrt{5}+1}} = \frac{5(\sqrt{5}+1)}{5+\sqrt{5}} = \sqrt{5}.$$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

La matrice di adiacenza per il 5-ciclo C_5

La rappresentazione ortonormale data da questa costruzione è, ovviamente, esattamente l'"ombrello di Lovász".

Ad esempio, per $m = 7$ tutto quello che sappiamo è

$$\sqrt[4]{108} \leq \Theta(C_7) \leq \frac{7}{1+(\cos\frac{\pi}{7})^{-1}},$$

che è $3.2237 \leq \Theta(C_7) \leq 3.3177$

Ma che ne è di C_7, C_9 e degli altri cicli dispari? Considerando $\alpha(C_m^2)$, $\alpha(C_m^3)$ ed altre piccole potenze, la minorazione $\frac{m-1}{2} \leq \Theta(C_m)$ può certamente essere aumentata, ma per nessun $m \geq 7$ dispari le minorazioni migliori si accordano con la maggiorazione data in (8). Quindi, vent'anni dopo la meravigliosa dimostrazione di Lovász che $\Theta(C_5) = \sqrt{5}$, questi problemi restano aperti e sono ritenuti molto difficili — ma dopo tutto abbiamo già avuto questa situazione in passato.

Gli autovalori di C_m

Si guardi la matrice di adiacenza A del ciclo C_m. Per trovare gli autovalori (e gli autovettori) usiamo le radici m-esime dell'unità. Queste sono date da $1, \zeta, \zeta^2, \ldots, \zeta^{m-1}$ per $\zeta = e^{\frac{2\pi i}{m}}$ — si veda il riquadro a pagina 29.

Sia $\lambda = \zeta^k$ una di queste radici, allora affermiamo che $(1, \lambda, \lambda^2, \ldots, \lambda^{m-1})^T$ è un autovettore di A per l'autovalore $\lambda+\lambda^{-1}$. In effetti, in base alla costruzione di A troviamo

$$A\begin{pmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{m-1} \end{pmatrix} = \begin{pmatrix} \lambda & + & \lambda^{m-1} \\ \lambda^2 & + & 1 \\ \lambda^3 & + & \lambda \\ & \vdots & \\ 1 & + & \lambda^{m-2} \end{pmatrix} = (\lambda+\lambda^{-1})\begin{pmatrix} 1 \\ \lambda \\ \lambda^2 \\ \vdots \\ \lambda^{m-1} \end{pmatrix}.$$

Poiché i vettori $(1, \lambda, \ldots, \lambda^{m-1})$ sono indipendenti (formano una cosiddetta matrice di Vandermonde) concludiamo che per m dispari

$$\begin{aligned} \zeta^k + \zeta^{-k} &= [(\cos(2k\pi/m) + i\sin(2k\pi/m)] \\ &\quad + [\cos(2k\pi/m) - i\sin(2k\pi/m)] \\ &= 2\cos(2k\pi/m) \qquad (0 \leq k \leq \tfrac{m-1}{2}) \end{aligned}$$

sono tutti gli autovalori di A. Ora il coseno è una funzione decrescente, e pertanto

$$2\cos\Big(\frac{(m-1)\pi}{m}\Big) = -2\cos\frac{\pi}{m}$$

è il più piccolo autovalore di A.

Bibliografia

[1] V. CHVÁTAL: *Linear Programming,* Freeman, New York 1983.

[2] W. HAEMERS: *Eigenvalue methods,* in: "Packing and Covering in Combinatorics" (A. Schrijver, ed.), Math. Centre Tracts **106** (1979), 15-38.

[3] L. LOVÁSZ: *On the Shannon capacity of a graph,* IEEE Trans. Information Theory **25** (1979), 1-7.

[4] C. E. SHANNON: *The zero-error capacity of a noisy channel,* IRE Trans. Information Theory **3** (1956), 3-15.

Di amici e politici

Capitolo 34

Non è noto chi abbia sollevato per primo il seguente problema né chi gli abbia fornito una connotazione umana. Eccolo:

> *Si supponga che in un gruppo di individui abbiamo la situazione in cui ogni coppia di persone abbia esattamente un amico comune. Allora vi è sempre un soggetto (il "politico") che è amico di tutti.*

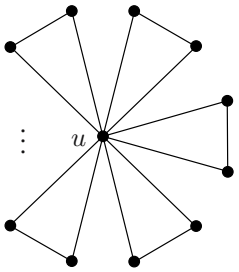
"Un sorriso da politico"

In gergo matematico questo teorema è chiamato il *teorema dell'amicizia*.

Prima di affrontare la dimostrazione riformuliamo il problema nei termini della teoria dei grafi. Consideriamo le persone come l'insieme dei vertici V e congiungiamo due vertici con uno arco se le persone corrispondenti sono amiche. Supponiamo tacitamente che l'amicizia sia sempre in due direzioni, ovvero, che se u sia amico di v, allora v sia a sua volta amico di u, ed inoltre che nessuno sia amico di se stesso. Pertanto il teorema assume la forma seguente:

Teorema. *Si supponga che G sia un grafo finito in cui ogni coppia di vertici ha esattamente un vicino in comune. Allora vi è un vertice adiacente a tutti gli altri vertici.*

Si noti che esistono grafi finiti con questa proprietà: si veda la figura, in cui u è il politico. Tuttavia, questi "grafi a mulinello" si rivelano anche essere gli unici grafi con la proprietà desiderata. In effetti, non è difficile verificare che in presenza di un politico sono possibili solo grafi a mulinello.

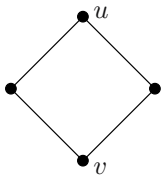


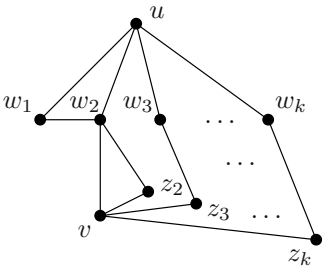
Un grafo a mulinello

Sorprendentemente, il teorema dell'amicizia non vale per grafi infiniti! Effettivamente, per costruire un controesempio per induzione si può cominciare ad esempio con un 5-ciclo ed aggiungere successivamente vicini comuni per tutte le coppie di vertici nel grafo che non ne hanno ancora. Questo porta ad un grafo dell'amicizia (calcolabilmente) infinito privo di politico.

Esistono svariate dimostrazioni del teorema dell'amicizia, ma la prima dimostrazione, fornita da Paul Erdős, Alfred Rényi e Vera Sós, è ancora la migliore.

■ **Dimostrazione.** Supponiamo che l'affermazione sia falsa e che G sia un controesempio, ovvero, che nessun vertice di G sia adiacente a tutti gli altri vertici. Per ottenere una contraddizione procediamo in due passi. La prima parte è basata sul calcolo combinatorio, la seconda sull'algebra lineare.





(1) Affermiamo che G è un grafo regolare, ovvero, $d(u) = d(v)$ per ogni $u, v \in V$.

Si noti innanzitutto che la condizione del teorema implica che non vi sono cicli di lunghezza 4 in G. Chiamiamo questa la *condizione* C_4.

Dimostriamo dapprima che ogni coppia di vertici *non adiacenti* u e v ha grado uguale $d(u) = d(v)$. Supponiamo che $d(u) = k$, dove $w_1, \ldots, w_k$ sono i vicini di u. Esattamente uno dei w_i, diciamo w_2, è adiacente a v, e w_2 è adiacente ad esattamente uno degli altri w_i, poniamo w_1, cosicché abbiamo la situazione della figura a lato. Il vertice v ha con w_1 il vicino comune w_2 e con w_i $(i \geq 2)$ un vicino comune z_i $(i \geq 2)$. In base alla condizione C_4, tutti questi z_i devono essere distinti. Concludiamo che $d(v) \geq k = d(u)$ e pertanto $d(u) = d(v) = k$ per simmetria.

Per terminare la dimostrazione di **(1)**, si osservi che ogni vertice diverso da w_2 non è adiacente né a u né a v e pertanto ha grado k, in base a quanto già dimostrato. Ma poiché w_2 ha anche un non-vicino, anch'esso ha grado k e pertanto G è k-regolare.

Sommando sui gradi dei k vicini di u otteniamo k^2. Poiché ogni vertice (tranne u) ha esattamente un vicino in comune con u, abbiamo contato ogni vertice una sola volta, eccetto u, che è stato contato k volte. Pertanto il numero totale dei vertici di G è

$$n \;=\; k^2 - k + 1. \tag{1}$$

(2) Il resto della dimostrazione è una splendida applicazione di alcuni risultati standard di algebra lineare. Si noti in primo luogo che k deve essere maggiore di 2, poiché per $k \leq 2$ solo $G = K_1$ e $G = K_3$ sono possibili in base a (1) ed entrambi sono grafi a mulinello banali. Si consideri la matrice di adiacenza $A = (a_{ij})$, definita a pagina 246. In base alla parte **(1)**, ogni riga ha esattamente k 1 ed, in base alla condizione del teorema, per ogni coppia di righe vi è esattamente una colonna in cui entrambe hanno un 1. Si noti inoltre che la diagonale principale consiste di 0. Pertanto abbiamo

$$A^2 \;=\; \begin{pmatrix} k & 1 & \ldots & 1 \\ 1 & k & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \ldots & 1 & k \end{pmatrix} \;=\; (k-1)\, I + J\,,$$

dove I è la matrice identità e J la matrice di tutti 1. È immediatamente verificabile che J ha come autovalori n (con molteplicità 1) e 0 (con molteplicità $n-1$). Ne segue che A^2 ha come autovalori $k - 1 + n = k^2$ (con molteplicità 1) e $k - 1$ (con molteplicità $n - 1$).

Poiché A è simmetrica e pertanto diagonalizzabile, concludiamo che A ha come autovalori k (con molteplicità 1) e $\pm\sqrt{k-1}$. Supponiamo che r degli autovalori siano uguali a $\sqrt{k-1}$ e che s di essi siano uguali a $-\sqrt{k-1}$, con $r + s = n - 1$. Adesso siamo quasi arrivati. Poiché la somma degli autovalori di A è uguale alla traccia (che vale 0), troviamo

$$k + r\sqrt{k-1} - s\sqrt{k-1} \;=\; 0,$$

ed in particolare, $r \neq s$, e

$$\sqrt{k-1} \;=\; \frac{k}{s-r}.$$

Ora se la radice quadrata $\sqrt{m}$ di un numero naturale m è razionale, allora m è un numero intero! Una dimostrazione elegante di questo fatto fu presentata da Dedekind nel 1858: sia n_0 il più piccolo numero naturale tale che $n_0\sqrt{m} \in \mathbb{N}$. Se $\sqrt{m} \notin \mathbb{N}$, allora esiste $\ell \in \mathbb{N}$ tale che $0 < \sqrt{m} - \ell < 1$. Ponendo $n_1 := n_0(\sqrt{m} - \ell)$, troviamo $n_1 \in \mathbb{N}$ e $n_1\sqrt{m} = n_0(\sqrt{m} - \ell)\sqrt{m} = n_0 m - \ell(n_0\sqrt{m}) \in \mathbb{N}$. Insieme a $n_1 < n_0$ ciò dà una contraddizione, tenuto conto della scelta di n_0.
Tornando alla nostra equazione, poniamo $h = \sqrt{k-1} \in \mathbb{N}$, allora

$$h(s-r) \;=\; k \;=\; h^2 + 1.$$

Poiché h divide $h^2 + 1$ e h^2, troviamo che h deve essere uguale a 1, dunque $k = 2$, cosa che abbiamo già escluso. Pertanto siamo giunti ad una contraddizione e la dimostrazione è completa. □

Tuttavia, la storia non è del tutto conclusa. Riformuliamo il nostro teorema nel modo seguente: supponiamo che G sia un grafo con la proprietà che tra ogni coppia di vertici esista esattamente un cammino di lunghezza 2. Chiaramente, questa è una formulazione equivalente del teorema dell'amicizia. Il nostro teorema afferma allora che gli unici grafi di questa forma sono grafi a mulinello. Ma cosa accade se consideriamo cammini di lunghezza maggiore di 2? Una congettura di Anton Kotzig afferma che la situazione analoga è impossibile.

Congettura di Kotzig. *Sia $\ell > 2$. Allora non vi sono grafi finiti con la proprietà che tra ogni coppia di vertici vi è esattamente un cammino di lunghezza ℓ.*

Kotzig stesso verificò la sua congettura per $\ell \leq 8$. In [3] la sua congettura è dimostrata fino a $\ell = 20$ ed A. Kostochka ci ha detto recentemente che essa è già verificata per ogni $\ell \leq 33$. Una dimostrazione generale, tuttavia, sembra lontana ...

Bibliografia

[1] P. ERDŐS, A. RÉNYI & V. SÓS: *On a problem of graph theory,* Studia Sci. Math. **1** (1966), 215-235.

[2] A. KOTZIG: *Regularly k-path connected graphs,* Congressus Numerantium **40** (1983), 137-141.

[3] A. KOSTOCHKA: *The nonexistence of certain generalized friendship graphs,* in: "Combinatorics" (Eger, 1987), Colloq. Math. Soc. János Bolyai **52**, North-Holland, Amsterdam 1988, 341-356.

Le probabilità semplificano (talvolta) il contare

Capitolo 35

Avendo iniziato questo libro con le prime pubblicazioni di Paul Erdős sulla teoria dei numeri, lo concludiamo discutendo quella che sarà probabilmente considerata la sua eredità più longeva — l'introduzione, insieme con Alfred Rényi, del *metodo probabilistico*. Formulato nel modo più semplice esso afferma:

> *Se, in un insieme dato di oggetti, la probabilità che un oggetto non abbia una certa proprietà è minore di* 1*, allora deve esistere un oggetto con questa proprietà.*

Pertanto abbiamo un risultato di *esistenza*. Può essere (e spesso è) molto difficile trovare tale oggetto, ma sappiamo che esiste. Presentiamo qui tre esempi (di sofisticatezza crescente) di questo metodo probabilistico dovuti a Erdős e terminiamo con un'applicazione recente particolarmente elegante.

Come riscaldamento, si consideri una famiglia $\mathcal{F}$ di sottoinsiemi A_i, tutti di dimensione $d \geq 2$, di un insieme di base finito X. Diciamo che $\mathcal{F}$ è 2-*colorabile* se esiste una colorazione di X con due colori tale che in ogni insieme A_i appaiano entrambi i colori. È evidente che le famiglie non possono essere tutte colorate in questo modo. Come esempio, si prendano *tutti* i sottoinsiemi di dimensione d di un $(2d-1)$-insieme X. Allora non importa come coloriamo X con due colori, devono esserci d elementi colorati in modo analogo. D'altro canto, è ugualmente chiaro che ogni sottofamiglia di una famiglia 2-colorabile di d-insiemi è essa stessa 2-colorabile. Siamo pertanto interessati al *più piccolo* numero $m = m(d)$ per cui esista una famiglia con m insiemi che non sia 2-colorabile. Formulato diversamente, $m(d)$ è il più grande numero che garantisce che *ogni* famiglia con meno di $m(d)$ insiemi sia 2-colorabile.

Teorema 1. *Ogni famiglia di al massimo* 2^{d-1} d*-insiemi è* 2*-colorabile, ovvero,* $m(d) > 2^{d-1}$.

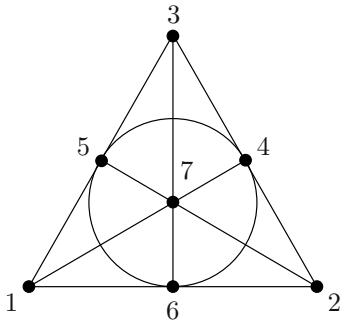
■ **Dimostrazione.** Supponiamo che $\mathcal{F}$ sia una famiglia di d-insiemi con al massimo 2^{d-1} insiemi. Coloriamo arbitrariamente X con due colori, in modo che tutte le colorazioni siano equiprobabili. Per ogni insieme $A \in \mathcal{F}$ sia E_A l'evento in cui tutti gli elementi di A ricevano lo stesso colore. Poiché vi sono esattamente due colorazioni di questo tipo, abbiamo

$$\text{Prob}(E_A) \;=\; \left(\tfrac{1}{2}\right)^{d-1},$$

e pertanto con $m = |\mathcal{F}| \leq 2^{d-1}$ (si noti che gli eventi E_A non sono disgiunti)

$$\text{Prob}(\bigcup_{A\in\mathcal{F}} E_A) \;<\; \sum_{A\in\mathcal{F}} \text{Prob}(E_A) \;=\; m\,(\tfrac{1}{2})^{d-1} \;\leq\; 1.$$

Concludiamo che esistono alcune 2-colorazioni di X senza un insieme d colorato con un solo colore da $\mathcal{F}$ e questa è proprio la nostra condizione di 2-colorabilità. □

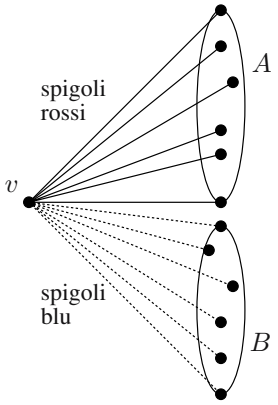


Una maggiorazione per $m(d)$, pari all'incirca a $d^2 2^d$, fu stabilita sempre da Erdős, ancora usando il metodo probabilistico, questa volta considerando insiemi arbitrari ed una colorazione fissa. Quanto ai valori esatti, sono noti solo i primi due, $m(2) = 3$, $m(3) = 7$. Ovviamente, $m(2) = 3$ è realizzato dal grafo K_3, mentre la configurazione di Fano dà $m(3) \leq 7$. Qui $\mathcal{F}$ consiste nei sette insiemi di 3 elementi della figura (compreso l'insieme circolare $\{4, 5, 6\}$). Il lettore potrà trovare divertente dimostrare che $\mathcal{F}$ necessita di 3 colori. Dimostrare che tutte le famiglie di 6 insiemi di 3 elementi sono 2-colorabili, e pertanto $m(3) = 7$, richiede un po' di attenzione in più.

Il nostro esempio successivo è il classico di questo campo — i numeri di Ramsey. Si consideri il grafo completo K_N su N vertici. Diciamo che K_N ha la proprietà (m, n) se, indipendentemente da come coloriamo gli archi di K_N di rosso e blu, esiste sempre un sottografo completo su m vertici con tutti gli archi colorati di rosso o un sottografo completo su n vertici con tutti gli archi colorati di blu. È chiaro che se K_N ha la proprietà (m, n), allora ciò vale anche per ogni K_s tale che $s \geq N$. Pertanto, come nel primo esempio, chiediamo il *più piccolo* numero N (se esiste) con questa proprietà — e questo è il *numero di Ramsey* $R(m, n)$.

Per cominciare, abbiamo certamente $R(m, 2) = m$ perché o tutti gli archi di K_m sono rossi o vi è un arco blu, il che risulta in un K_2 blu. Per simmetria, abbiamo $R(2, n) = n$. Ora, supponiamo che $R(m-1, n)$ e $R(m, n-1)$ esistano. Dimostriamo allora che $R(m, n)$ esiste e che

$$R(m, n) \;\leq\; R(m-1, n) \;+\; R(m, n-1). \tag{1}$$



Supponiamo che $N = R(m-1, n) + R(m, n-1)$ e consideriamo una colorazione arbitraria rossa e blu di K_N. Per un vertice dato v, sia A l'insieme dei vertici collegati con v da un arco rosso, e B quello dei vertici collegati da un arco blu.

Poiché $|A| + |B| = N - 1$, troviamo che o $|A| \geq R(m-1, n)$ o $|B| \geq R(m, n-1)$. Supponiamo che $|A| \geq R(m-1, n)$, l'altro caso essendo analogo. Allora per definizione di $R(m-1, n)$, in A esiste o un sottoinsieme A_R di dimensione $m-1$ i cui archi sono tutti colorati di rosso il quale insieme a v dà un K_m rosso, oppure un sottoinsieme A_B di dimensione n con tutti gli archi colorati di blu. Inferiamo che K_N soddisfa la proprietà (m, n) e ne segue la disuguaglianza (1).

Combinando (1) con i valori iniziali $R(m, 2) = m$ e $R(2, n) = n$, ottenia-

mo dalla famigliare formula ricorsiva per i coefficienti binomiali

$$R(m,n) \;\leq\; \binom{m+n-2}{m-1},$$

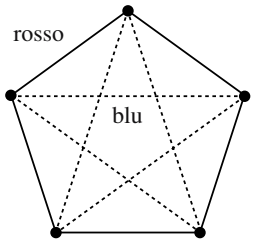
ed in particolare

$$R(k,k) \;\leq\; \binom{2k-2}{k-1} \;=\; \binom{2k-3}{k-1} + \binom{2k-3}{k-2} \;\leq\; 2^{2k-3}. \quad (2)$$

Quello che ci interessa veramente ora è una minorazione per $R(k,k)$. Ciò si riduce al dimostrare per un $N < R(k,k)$ il più grande possibile che *esiste* una colorazione degli archi tale che non risulti alcun K_k rosso o blu. Ed è qui che il metodo probabilistico entra in gioco.

Teorema 2. *Per ogni $k \geq 2$, vale la seguente minorazione per i numeri di Ramsey:*

$$R(k,k) \;\geq\; 2^{\frac{k}{2}}.$$



■ **Dimostrazione.** Abbiamo $R(2,2) = 2$. Da (2) sappiamo che $R(3,3) \leq 6$ ed il pentagono colorato come nella figura mostra che $R(3,3) = 6$.

Supponiamo ora che $k \geq 4$. Supponiamo che $N < 2^{\frac{k}{2}}$ e consideriamo tutte le colorazioni rosso-blu dove coloriamo ogni arco indipendentemente di rosso o blu con probabilità $\frac{1}{2}$. Pertanto tutte le colorazioni sono ugualmente probabili con probabilità $2^{-\binom{N}{2}}$. Sia A un insieme di vertici di dimensione k. La probabilità dell'evento A_R che gli archi in A siano tutti colorati di rosso è allora $2^{-\binom{k}{2}}$. Pertanto segue che la probabilità p_R che *un* k-insieme sia colorato interamente di rosso è maggiorata da

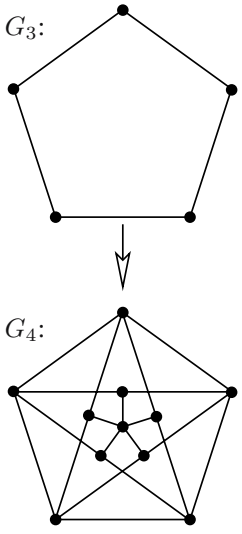
$$p_R \;=\; \text{Prob}\Big(\bigcup_{|A|=k} A_R\Big) \;\leq\; \sum_{|A|=k} \text{Prob}(A_R) \;=\; \binom{N}{k} 2^{-\binom{k}{2}}.$$

Ora con $N < 2^{\frac{k}{2}}$ e $k \geq 4$, usando $\binom{N}{k} \leq \frac{N^k}{2^{k-1}}$ per $k \geq 2$ (si veda a pagina 12), abbiamo

$$\binom{N}{k} 2^{-\binom{k}{2}} \;\leq\; \frac{N^k}{2^{k-1}} 2^{-\binom{k}{2}} \;<\; 2^{\frac{k^2}{2} - \binom{k}{2} - k + 1} \;=\; 2^{-\frac{k}{2}+1} \leq \frac{1}{2}.$$

Pertanto $p_R < \frac{1}{2}$ e, per simmetria, $p_B < \frac{1}{2}$ per la probabilità che k vertici con tutti gli archi fra loro siano colorati di blu. Concludiamo che $p_R + p_B < 1$ per $N < 2^{\frac{k}{2}}$, quindi *deve* esistere una colorazione senza K_k rossi o blu, il che significa che K_N non ha la proprietà (k,k). □

Ovviamente, vi è una discreta distanza tra la minorazione e la maggiorazione per $R(k,k)$. Tuttavia, malgrado la semplicità di questa *Book Proof*, non è stata trovata una minorazione con un esponente migliore per un k generico negli oltre 50 anni successivi al risultato di Erdős. In effetti, nessuno è stato in grado di dimostrare una minorazione della forma $R(k,k) > 2^{(\frac{1}{2}+\varepsilon)k}$ né una maggiorazione della forma $R(k,k) < 2^{(2-\varepsilon)k}$ per un $\varepsilon > 0$ fissato.



Costruzione del grafico di Mycielski

Il nostro terzo risultato è un'altra splendida illustrazione del metodo probabilistico. Si considerino un grafo G su n vertici ed il suo numero cromatico $\chi(G)$. Se $\chi(G)$ è alto, ovvero, se abbiamo bisogno di molti colori, allora potremmo sospettare che G contenga un grande sottografo completo. Tuttavia, ciò è distante dalla realtà. Già negli anni $'40$ Blanche Descartes costruiva grafi con un numero cromatico arbitrariamente grande e senza triangoli ovvero con ogni ciclo di lunghezza minima 4 e così fecero molti altri (si veda il riquadro qui sotto).

Tuttavia, in questi esempi vi erano molti cicli di lunghezza 4. Possiamo fare ancora meglio? Possiamo stabilire che non vi sono cicli di lunghezza breve ed avere comunque un numero cromatico arbitrariamente alto? Sì, possiamo! Per precisare la situazione, indichiamo con $\gamma(G)$ la *lunghezza di un ciclo minimo* [NdT: *girth* nell'edizione inglese] in G; allora abbiamo il seguente teorema, dimostrato per la prima volta da Paul Erdős.

Grafi senza triangoli con numero cromatico alto

Ecco una successione di grafi senza triangoli $G_3, G_4, \ldots$ con

$$\chi(G_n) \;=\; n.$$

Cominciando con $G_3 = C_5$, il 5-ciclo; pertanto $\chi(G_3) = 3$. Supponiamo di aver già costruito G_n sull'insieme dei vertici V. Il nuovo grafo G_{n+1} ha come insieme dei vertici $V \cup V' \cup \{z\}$, dove i vertici $v' \in V'$ corrispondono biiettivamente a $v \in V$ e z è un altro vertice singolo. Gli archi di G_{n+1} cadono in tre classi: prima, prendiamo tutti gli archi di G_n; in secondo luogo ogni vertice v' è collegato precisamente con i vicini di v in G_n; in terzo luogo z è collegato con tutti i $v' \in V'$. Pertanto da $G_3 = C_5$ otteniamo come G_4 il cosiddetto *grafo di Mycielski*.

Chiaramente, G_{n+1} è ancora senza triangoli. Per dimostrare $\chi(G_{n+1}) = n + 1$ usiamo l'induzione su n. Si prenda una qualsiasi colorazione con n colori di G_n e si consideri una classe di colori C. Deve esistere un vertice $v \in C$ adiacente ad almeno un vertice di ognuna delle altre classi di colore; altrimenti potremmo distribuire i vertici di C sulle altre $n - 1$ classi di colori, il che risulterebbe in $\chi(G_n) \leq n - 1$. Ma ora è chiaro che v' (il vertice in V' corrispondente a v) deve ricevere lo stesso colore di v in questa colorazione a n colori. Quindi, tutti gli n colori appaiono in V' ed abbiamo bisogno di un nuovo colore per z.

Teorema 3. *Per ogni $k \geq 2$, esiste un grafo G con numero cromatico $\chi(G) > k$ e una lunghezza del ciclo minimo $\gamma(G) > k$.*

La strategia è simile a quella delle dimostrazioni precedenti: consideriamo un certo spazio di probabilità sui grafi e procediamo mostrando che la probabilità che $\chi(G) \leq k$ è inferiore a $\frac{1}{2}$ ed analogamente la probabilità che

$\gamma(G) \leq k$ è inferiore a $\frac{1}{2}$. Conseguentemente, deve esistere un grafo con le proprietà desiderate.

■ **Dimostrazione.** Sia $V = \{v_1, v_2, \ldots, v_n\}$ l'insieme dei vertici e p un numero fisso tra 0 e 1, da scegliere attentamente più tardi. Il nostro spazio di probabilità $\mathcal{G}(n,p)$ consiste di tutti i grafi su V in cui i singoli archi appaiano con probabilità p, indipendentemente l'uno dall'altro. In altre parole, stiamo parlando di un esperimento di Bernoulli in cui tiriamo ogni arco con probabilità p. Come esempio, la probabilità Prob(K_n) per il grafo completo è Prob$(K_n) = p^{\binom{n}{2}}$. In generale, abbiamo Prob$(H) = p^m(1-p)^{\binom{n}{2}-m}$ se il grafo H su V ha esattamente m archi.

Consideriamo dapprima il numero cromatico $\chi(G)$. Con $\alpha = \alpha(G)$ indichiamo il *numero d'indipendenza* ovvero la dimensione del più grande insieme indipendente in G. Poiché in una colorazione con $\chi = \chi(G)$ colori tutte le classi di colori sono indipendenti (e pertanto di dimensione $\leq \alpha$), inferiamo $\chi\alpha \geq n$. Pertanto se α è piccolo in confronto a n, allora χ deve essere grande, che è ciò che vogliamo.

Supponiamo che $2 \leq r \leq n$. La probabilità che un insieme fissato di dimensione r in V sia indipendente è $(1-p)^{\binom{r}{2}}$ e concludiamo con lo stesso argomento del Teorema 2 che

$$\begin{aligned}\text{Prob}(\alpha \geq r) &\leq \binom{n}{r}(1-p)^{\binom{r}{2}} \\ &\leq n^r(1-p)^{\binom{r}{2}} = (n(1-p)^{\frac{r-1}{2}})^r \leq (ne^{-p(r-1)/2})^r,\end{aligned}$$

poiché $1 - p \leq e^{-p}$ per ogni p.

Dato un qualunque $k > 0$ fisso scegliamo ora $p := n^{-\frac{k}{k+1}}$ e procediamo per mostrare che per n abbastanza grande,

$$\text{Prob}\Big(\alpha \geq \frac{n}{2k}\Big) < \frac{1}{2}. \tag{3}$$

Effettivamente, poiché $n^{\frac{1}{k+1}}$ cresce più rapidamente di $\log n$, abbiamo $n^{\frac{1}{k+1}} \geq 6k \log n$ per n grandi abbastanza e pertanto $p \geq 6k\frac{\log n}{n}$. Per $r := \lceil \frac{n}{2k} \rceil$ questo dà $pr \geq 3 \log n$ e dunque

$$ne^{-p(r-1)/2} = ne^{-\frac{pr}{2}}e^{\frac{p}{2}} \leq ne^{-\frac{3}{2}\log n}e^{\frac{1}{2}} = n^{-\frac{1}{2}}e^{\frac{1}{2}} = (\tfrac{e}{n})^{\frac{1}{2}},$$

che converge a 0 per n che tende all'infinito. Pertanto (3) vale per ogni $n \geq n_1$.

Ora esaminiamo il secondo parametro, $\gamma(G)$. Per il k dato vogliamo dimostrare che non vi sono troppi cicli di lunghezza $\leq k$. Sia i tra 3 e k e $A \subseteq V$ un insieme fisso di dimensione i. Il numero di i-cicli possibili su A è chiaramente il numero di permutazioni cicliche di A diviso per 2 (dal momento che possiamo percorrere il ciclo in entrambe le direzioni) e quindi uguale a $\frac{(i-1)!}{2}$. Il numero totale di i-cicli possibili è pertanto $\binom{n}{i}\frac{(i-1)!}{2}$ ed ognuno di tali cicli C appare con probabilità p^i. Sia X la variabile aleatoria che

conta il numero di cicli di lunghezza $\leq k$. Al fine di stimare X usiamo due strumenti semplici ma splendidi. Il primo è la linearità del valore atteso ed il secondo è la disuguaglianza di Markov per le variabili aleatorie non negative, che afferma

$$\text{Prob}(X \geq a) \ \leq \ \frac{EX}{a},$$

dove EX è il valore atteso di X. Si veda l'appendice al Capitolo 14 per entrambi gli strumenti.

Sia X_C la varabile aleatoria indicatrice del ciclo C di, poniamo, lunghezza i. In sostanza, poniamo $X_C = 1$ o 0 a seconda che C appaia nel grafo o meno; pertanto $EX_C = p^i$. Poiché X conta il numero di tutti i cicli di lunghezza $\leq k$ abbiamo $X = \sum X_C$ e dunque per linearità

$$EX \ = \ \sum_{i=3}^{k} \binom{n}{i} \frac{(i-1)!}{2} p^i \ \leq \ \frac{1}{2} \sum_{i=3}^{k} n^i p^i \ \leq \ \frac{1}{2}(k-2) n^k p^k$$

dove l'ultima disuguaglianza vale per il fatto che $np = n^{\frac{1}{k+1}} \geq 1$. Applicando ora la disuguaglianza di Markov con $a = \frac{n}{2}$, otteniamo

$$\text{Prob}(X \geq \tfrac{n}{2}) \ \leq \ \frac{EX}{n/2} \ \leq \ (k-2)\frac{(np)^k}{n} \ = \ (k-2) n^{-\frac{1}{k+1}}.$$

Poiché il membro destro tende a 0 con n che tende all'infinito, inferiamo che $p(X \geq \frac{n}{2}) < \frac{1}{2}$ per $n \geq n_2$.

Ci siamo quasi. La nostra analisi ci dice che per $n \geq \max(n_1, n_2)$ esiste un grafo H su n vertici con $\alpha(H) < \frac{n}{2k}$ e meno di $\frac{n}{2}$ cicli di lunghezza $\leq k$. Si elimini un vertice da ognuno di questi cicli e sia G il grafo risultante. Allora $\gamma(G) > k$ vale in ogni caso. Poiché G contiene più di $\frac{n}{2}$ vertici e soddisfa $\alpha(G) \leq \alpha(H) < \frac{n}{2k}$, troviamo

$$\chi(G) \ \geq \ \frac{n/2}{\alpha(G)} \ \geq \ \frac{n}{2\alpha(H)} \ > \ \frac{n}{n/k} \ = \ k,$$

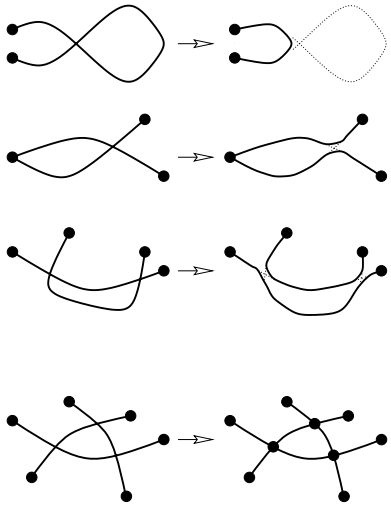
e la dimostrazione è terminata. □

Sono note delle costruzioni esplicite di grafi (di dimensioni enormi) con lunghezza del ciclo minimo e numero cromatico alti. (Al contrario, non si sa come costruire colorazioni rosso/blu senza clique monocromatiche grandi, la cui esistenza è data dal Teorema 2.) È davvero notevole che la dimostrazione di Erdős provi l'esistenza di grafi relativamente piccoli con numero cromatico e lunghezza del ciclo minimo alti.

Per terminare la nostra incursione nel mondo probabilistico discutiamo un risultato importante nella teoria geometrica dei grafi (che di nuovo risale a Paul Erdős) la cui stupefacente *Book Proof* è molto recente.

Si consideri un grafo semplice $G = G(V, E)$ con n vertici e m archi. Vogliamo inserire G nel piano, proprio come abbiamo fatto per i grafi planari.

Ora, sappiamo dal Capitolo 11 — come conseguenza della formula di Eulero — che un grafo planare semplice G ha al massimo $3n-6$ archi. Pertanto se m è maggiore di $3n-6$, deve esserci incrocio di archi. Il *numero di incroci* $\mathrm{cr}(G)$ è allora definito naturalmente: esso è il minor numero di incroci fra tutti i disegni di G, dove gli incroci di più di due archi in un punto non sono consentiti. Pertanto $\mathrm{cr}(G) = 0$ se e solo se G è planare.



In un tale disegno minimo sono escluse le tre situazioni seguenti:

- Nessun arco può incrociare se stesso.
- Gli archi con un estremo comune non possono incrociarsi.
- Nessuna coppia di archi può incrociarsi due volte.

Questo perché in tutti i casi citati possiamo costruire un disegno diverso dello stesso grafo con meno incroci, usando le operazioni indicate nella figura. Dunque, d'ora in poi supporremo che ogni disegno segua tali regole. Supponiamo che G sia disegnato nel piano con $\mathrm{cr}(G)$ incroci. Possiamo immediatamente derivare un limite inferiore sul numero degli incroci. Si consideri il seguente grafo H: i vertici di H sono quelli di G insieme con tutti i punti di incrocio e gli archi sono tutti parti degli archi originali mentre procediamo da punto d'incrocio a punto d'incrocio.

Il nuovo grafo H è ora piano e semplice (questo segue dalle nostre tre ipotesi!). Il numero dei vertici in H è $n+\mathrm{cr}(G)$ ed il numerodi archi è $m+2\mathrm{cr}(G)$, poiché ogni nuovo vertice ha grado 4. Invocando la maggiorazione sul numero di archi per i grafi piani troviamo pertanto

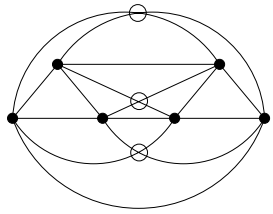
$$m + 2\,\mathrm{cr}(G) \;\leq\; 3(n + \mathrm{cr}(G)) - 6,$$

ovvero,

$$\mathrm{cr}(G) \;\geq\; m - 3n + 6. \tag{4}$$

Come esempio, per il grafo completo K_6 calcoliamo

$$\mathrm{cr}(K_6) \geq 15 - 18 + 6 = 3$$



ed in effetti esiste un grafo con solo tre incroci.

La limitazione (4) è abbastanza buona quando m è lineare in n, ma quando m è più grande rispetto a n, allora il quadro cambia e questo è il nostro teorema.

Teorema 4. *Sia G un grafo semplice con n vertici e m archi, dove $m \geq 4n$. Allora*

$$\mathrm{cr}(G) \;\geq\; \frac{1}{64}\frac{m^3}{n^2}.$$

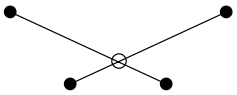
La storia di questo risultato, chiamato *lemma degli incroci*, è piuttosto interessante. Esso fu congetturato da Erdős e Guy nel 1973 (con $\frac{1}{64}$ sostituito da una costante c). Le prime dimostrazioni furono date da Leighton nel 1982 (con $\frac{1}{100}$ invece di $\frac{1}{64}$) ed indipendentemente da Ajtai, Chvátal, Newborn e Szemerédi. Il lemma degli incroci era appena conosciuto (in effetti, molte persone lo ritenevano ancora una congettura anche molto tempo dopo

che era stato effettivamente dimostrato), finché László Székely ne dimostrò l'utilità in una splendida pubblicazione, applicandolo ad una varietà di problemi geometrici estremali fino a quel momento difficili. La dimostrazione che ora presentiamo deriva da conversazioni via e-mail tra Bernard Chazelle, Micha Sharir ed Emo Welzl ed appartiene senza dubbio al *Libro*.

■ **Dimostrazione.** Si consideri un disegno minimale di G e sia p un numero tra 0 ed 1 (da scegliersi più tardi). Ora generiamo un sottografo di G, selezionando i vertici di G affinché giacciano nel sottografo con probabilità p, indipendentemente l'uno dall'altro. Il sottografo indotto che otteniamo in tal modo sarà chiamato G_p.

Siano n_p, m_p, X_p le variabili aleatorie che contano il numero dei vertici, degli archi e degli incroci in G_p. Poiché $\mathrm{cr}(G) - m + 3n \geq 0$ vale in base a (4) per *ogni* grafo, abbiamo certamente

$$E(X_p - m_p + 3n_p) \;\geq\; 0.$$



Procediamo ora per calcolare i singoli valori attesi $E(n_p)$, $E(m_p)$ e $E(X_p)$. Chiaramente, $E(n_p) = pn$ e $E(m_p) = p^2 m$, poiché un arco appare in G_p se e solo se entrambi i suoi estremi vi appaiono. Ed infine, $E(X_p) = p^4\mathrm{cr}(G)$, dal momento che un incrocio è presente in G_p se e solo se sono presenti tutti e quattro i vertici (distinti!) coinvolti.

Per la linearità del valore atteso troviamo pertanto

$$0 \;\leq\; E(X_p) \;-\; E(m_p) \;+\; 3E(n_p) \;=\; p^4\mathrm{cr}(G) - p^2 m + 3pn,$$

che è

$$\mathrm{cr}(G) \;\geq\; \frac{p^2 m - 3pn}{p^4} \;=\; \frac{m}{p^2} - \frac{3n}{p^3}. \tag{5}$$

Eccoci alla battuta finale: si ponga $p := \frac{4n}{m}$ (che vale al massimo 1 in base alla nostra supposizione), allora (5) diventa

$$\mathrm{cr}(G) \;\geq\; \frac{1}{64}\left[\frac{4m}{(n/m)^2} - \frac{3n}{(n/m)^3}\right] \;=\; \frac{1}{64}\frac{m^3}{n^2},$$

ed ecco fatto. □

Paul Erdős avrebbe adorato vedere questa dimostrazione.

Bibliografia

[1] M. AJTAI, V. CHVÁTAL, M. NEWBORN & E. SZEMERÉDI: *Crossing-free subgraphs,* Annals of Discrete Math. **12** (1982), 9-12.

[2] N. ALON & J. SPENCER: *The Probabilistic Method,* Second edition, Wiley-Interscience 2000.

[3] P. ERDŐS: *Some remarks on the theory of graphs,* Bulletin Amer. Math. Soc. **53** (1947), 292-294.

[4] P. ERDŐS: *Graph theory and probability,* Canadian J. Math. **11** (1959), 34-38.

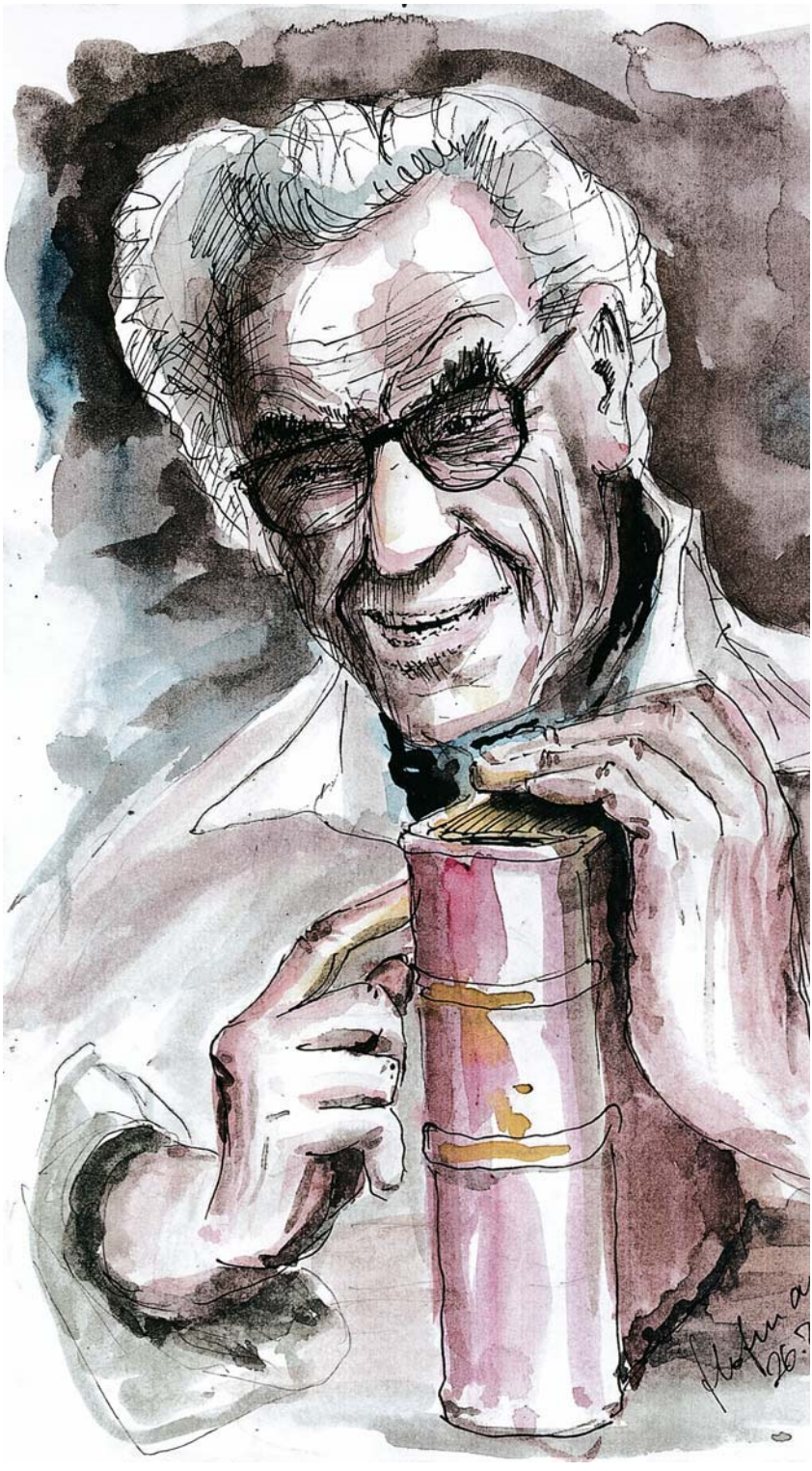
[5] P. ERDŐS: *On a combinatorial problem I,* Nordisk Math. Tidskrift **11** (1963), 5-10.

[6] P. ERDŐS & R. K. GUY: *Crossing number problems,* Amer. Math. Monthly **80** (1973), 52-58.

[7] P. ERDŐS & A. RÉNYI: *On the evolution of random graphs,* Magyar Tud. Akad. Mat. Kut. Int. Közl. **5** (1960), 17-61.

[8] T. LEIGHTON: *Complexity Issues in VLSI,* MIT Press, Cambridge MA 1983.

[9] L. A. SZÉKELY: *Crossing numbers and hard Erdős problems in discrete geometry,* Combinatorics, Probability, and Computing **6** (1997), 353-358.



A proposito delle illustrazioni

Siamo lieti di avere la possibilità ed il privilegio di illustrare questo volume con degli splendidi disegni originali di Karl Heinrich Hofmann (Darmstadt). Grazie!

I poliedri regolari a pagina 72 e la mappa pieghevole della sfera flessibile a pagina 82 sono di WAF Ruppert (Vienna).

Jürgen Richter-Gebert ha fornito due illustrazioni a pagina 74.

A pagina 229 è rappresentato il Weisman Art Museum di Minneapolis progettato da Frank Gehry. La fotografia della facciata ovest è di Chris Faust. La planimetria è quella della Dolly Fiterman Riverview Gallery, che si trova dietro la facciata ovest.

I ritratti di Bertrand, Cantor, Erdős, Eulero, Fermat, Herglotz, Hermite, Hilbert, Pólya, Littlewood e Sylvester provengono dall'archivio fotografico del Mathematisches Forschungsinstitut Oberwolfach, riprodotti su autorizzazione. (Molte grazie ad Annette Disch!)

I francobolli ritratto di Buffon, Chebyshev, Eulero e Ramanujan vengono dal sito di francobolli di matematica di Jeff Miller: http://jeff560.tripod.com grazie al suo generoso permesso.

L'immagine di Hermite è tratta dal primo volume della raccolta delle sue opere.

La fotografia di Claude Shannon è stata fornita dal MIT Museum ed è qui riprodotta con il loro permesso.

Il ritratto di Cayley è tratto dal "Photoalbum für Weierstraß" (edito da Reinhard Bölling, Vieweg 1994), col permesso dalla Kunstbibliothek, Staatliche Museen di Berlino, Preussischer Kulturbesitz.

Il ritratto di Cauchy è riprodotto col permesso delle Collections de l'École Polytechnique, Parigi.

L'immagine di Fermat è riprodotta dal libro di Stefan Hildebrandt e Anthony Tromba: *The Parsimonious Universe. Shape and Form in the Natural World*, Springer-Verlag, New York 1996.

Il ritratto di Ernst Witt viene dal volume 426 (1992) del Journal für die Reine und Angewandte Mathematik, col permesso di Walter de Gruyter Publishers. Esso fu realizzato intorno al 1941.

La fotografia di Karol Borsuk fu scattata nel 1967 da Isaac Namioka ed è riprodotta col suo gentile consenso.

Ringraziamo il Dr. Peter Sperner (Braunschweig) per il ritratto di suo padre, e Vera Sós per la foto di Paul Turán.

Grazie infine a Noga Alon per la foto ritratto di A. Nilli!

Indice analitico