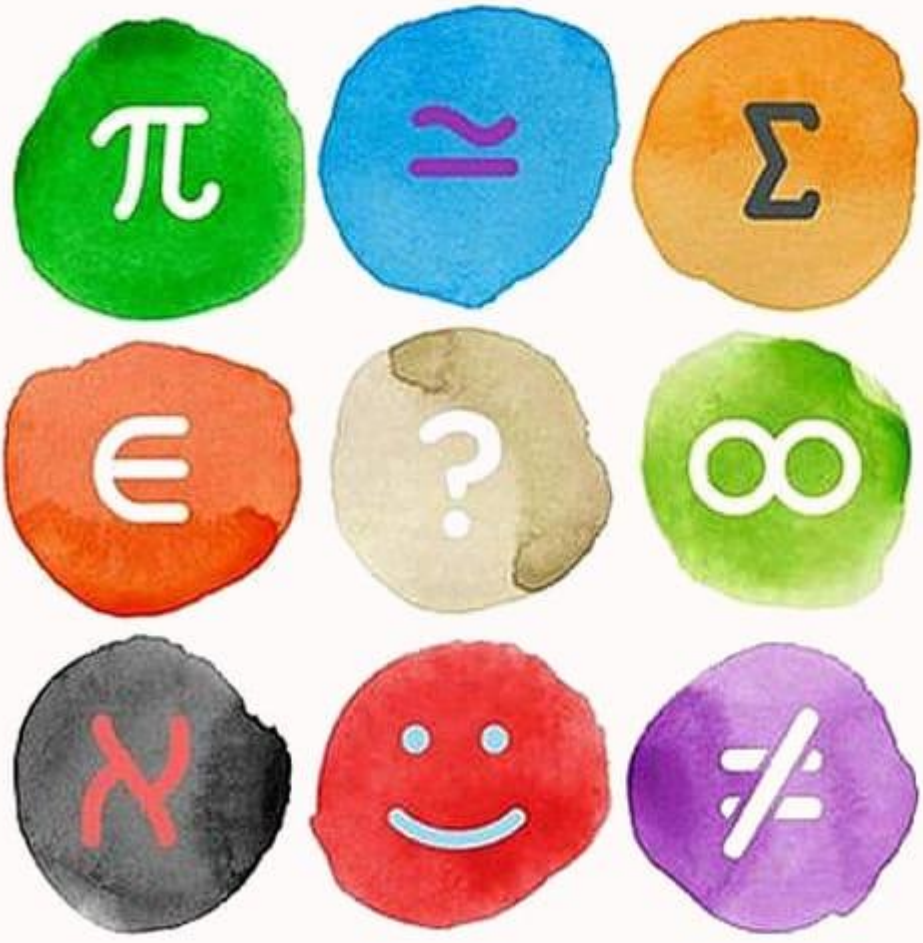
Ian Stewart

Los grandes problemas matemáticos



Reseña

Hay algunos problemas matemáticos cuya importancia va más allá de lo ordinario - como último teorema de Fermat o la conjetura de Goldbach - son los enigmas que definen las matemáticas. Los grandes problemas matemáticosexplica cuáles son estos problemas, por qué son importantes, lo que impulsa a los matemáticos a retos increíbles para resolverlos y cuál es su situación en el contexto de las matemáticas y las ciencias en general. A lo largo de la obra encontraremos problemas resueltos -como la Conjetura de Poincaré, resuelto por el genio excéntrico Grigori Perelman, que rechazó los honores académicos y un premio del millón de dólares por su trabajo- y otros que, como la Hipótesis de Riemann, siendo desconcertante después de siglos. Stewart es el guía ideal en este mundo misterioso y emocionante, que muestra cómo los matemáticos modernos se enfrentan a los retos establecidos por sus predecesores, ya que los grandes problemas matemáticos del pasado sucumben, a veces, a las nuevas técnicas e ideas del presente.

Índice

Prefacio

1. Grandes problemas
2. Territorio primo
3. El rompecabezas de pi
4. Cartografiando misterios
5. Simetría esférica
6. Nuevas soluciones para lo viejo
7. Márgenes estrechos
8. Caos orbital
9. Pautas en los primos
10. ¿Qué forma tiene una esfera?
11. No todos pueden ser fáciles
12. Pensamiento fluido
13. Enigma cuántico
14. Sueños diofánticos
15. Ciclos complejos
16. ¿Qué viene ahora?
17. Doce para el futuro

Glosario

Lecturas adicionales

Debemos saber, sabremos.
David Hilbert
Conferencia sobre problemas matemáticos con ocasión de su nombramiento como ciudadano honorario de Königsberg, en 1930[1]

Prefacio

El dominio de las matemáticas es inmenso, y está en continuo crecimiento y en continuo cambio. De entre las innumerables preguntas que plantean los matemáticos, y en su mayoría responden, algunas destacan como cimas imponentes que se alzan entre modestas estribaciones. Estas son las preguntas realmente grandes, los problemas difíciles y desafiantes que cualquier matemático daría su brazo derecho por resolver. Algunos permanecieron sin respuesta durante décadas, otros durante siglos, unos pocos durante milenios. Algunos aún tienen que ser vencidos. El último teorema de Fermat fue un enigma durante 350 años hasta que Andrew Wiles lo resolvió tras siete años de duro trabajo. La conjetura de Poincaré permaneció abierta durante más de cien años hasta que fue resuelta por el genio excéntrico Grigori Perelman, quien rechazó todos los honores académicos y un premio de un millón de dólares por su trabajo. La hipótesis de Riemann continúa

[1] El original alemán es: «Wir müssen wissen. Wir werden wissen». Se encuentra en una charla que grabó Hilbert para la radio. Ver Constante Reid, *Hilbert*, Springer, Berlín, 1970, p. 196

desconcertando a matemáticos de todo el mundo, tan impenetrable como siempre después de 150 años.

Los grandes problemas matemáticos contiene una selección de las preguntas más grandes que han impulsado la empresa matemática en direcciones radicalmente nuevas. Describe sus orígenes, explica por qué son importantes y los sitúa en el contexto de las matemáticas y la ciencia en general. Incluye problemas, tanto resueltos como no resueltos, que cubren más de dos mil años de desarrollo matemático, pero su foco se centra en cuestiones que o bien siguen hoy abiertas o bien han sido resueltas en los cincuenta últimos años.

Un objetivo básico de las matemáticas es revelar la simplicidad que subyace en preguntas aparentemente complicadas. Quizá esto no sea siempre evidente porque la idea que tiene el matemático de lo que es «simple» está basada en muchos conceptos técnicos y difíciles. Un aspecto importante de este libro es resaltar las profundas simplicidades, y evitar —o cuando menos explicar en términos sencillos— las complejidades.

* * * *

Las matemáticas son más novedosas, y más diversas, que lo que la mayoría de nosotros imaginamos. En una cruda estimación, el número de matemáticos que investigan en todo el mundo es de unos cien mil, y ellos producen más de *dos millones* de páginas de matemáticas nuevas cada año. No «números nuevos», que no es de lo que realmente trata la empresa. No «sumas nuevas» como las existentes, sino más grandes, aunque calculamos alguna sumas

muy grandes. Un reciente trabajo de álgebra, realizado por un equipo de unos veinticinco matemáticos, fue descrito como «un cálculo del tamaño de Manhattan».[i]

* * * *

Esto no era totalmente cierto; más bien pecaba de conservador. Lo que era del tamaño de Manhattan era la *respuesta*; el cálculo era mucho mayor. Esto es impresionante, pero lo que importa es la calidad, no la cantidad. El cálculo del tamaño de Manhattan es destacable en ambos aspectos, porque proporciona una información básica y muy valiosa sobre un grupo de simetría que parece ser importante en física cuántica, y es decididamente importante en matemáticas. Las matemáticas brillantes pueden ocupar una línea o pueden ocupar una enciclopedia, lo que el problema demande.

Cuando pensamos en matemáticas, lo que viene a la mente son páginas interminables llenas de fórmulas y símbolos. Sin embargo, esos dos millones de páginas contienen en general más palabras que símbolos. Las palabras están allí para explicar los antecedentes del problema, el curso del argumento, el significado de los cálculos y cómo todo encaja en el siempre creciente edificio de las matemáticas. Como comentaba el gran Carl Friedrich Gauss hacia 1800, la esencia de las matemáticas es «nociones, no notaciones». Ideas, no símbolos. Aun así, el lenguaje habitual para expresar ideas matemáticas es un lenguaje simbólico. Muchos artículos de investigación publicados contienen más símbolos que palabras. Las fórmulas tienen una precisión que no siempre las palabras pueden igualar.

Sin embargo, es a veces posible explicar las ideas aun prescindiendo de muchos de los símbolos. Este es el principio que guía *Los grandes problemas matemáticos*. Ilustra lo que hacen los matemáticos, cómo piensan y por qué su tema de estudio es interesante e importante. De forma significativa, muestra cómo los matemáticos de hoy están abordando los retos que dejaron sus predecesores, a medida que uno a uno los grandes enigmas del pasado se rinden a las poderosas técnicas del presente, que cambian las matemáticas y la ciencia del futuro. Las matemáticas figuran entre los más grandes logros de la humanidad, y sus grandes problemas —resueltos y no resueltos— han guiado y estimulado su sorprendente poder durante milenios, tanto pasados como por venir.

Coventry, junio de 2012

Capítulo 1
Grandes problemas

Los programas de televisión que tratan de matemáticas son raros, y los buenos son aún más raros. Uno de los mejores, tanto por la implicación e interés de la audiencia como por sus contenidos, trataba del último teorema de Fermat. El programa fue producido en 1996 por John Lynch para *Horizon*, la serie bandera de divulgación científica de la British Broadcasting Corporation. Simon Singh, quien también estuvo implicado en la realización, convirtió la historia en un libro que fue un éxito de ventas espectacular[2]. En una página web, Simon Singh señalaba que el impresionante éxito del programa fue una verdadera sorpresa:

> *Fueron cincuenta minutos de matemáticos hablando de matemáticas, que no es precisamente una receta obvia para un éxito televisivo, pero el resultado fue un programa que captó la imaginación del público y fue aclamado por la crítica. El programa ganó el BAFTA al mejor documental, un premio Italia, otros premios internacionales y una nominación a los Emmy. Esto prueba que las matemáticas pueden ser tan excitantes y apasionantes como cualquier otra materia sobre el planeta.*

Creo que hay varias razones para el éxito tanto del programa de televisión como del libro, y tienen implicaciones para las historias

[2] Simon Singh, *Fermat's Last Theorem*, Fourth Estate, 1997.

que quiero contar aquí. Para centrar la discusión, me concentraré en el documental de la televisión.

El último teorema de Fermat es uno de los problemas matemáticos verdaderamente grandes. Surge de un comentario en apariencia inocuo que uno de los matemáticos más importantes del siglo XVII escribió en el margen de un libro de texto clásico. El problema se hizo tristemente famoso porque nadie pudo demostrar lo que afirmaba la nota que Pierre de Fermat escribió en el margen, y así estuvieron las cosas durante más de trescientos años pese a los enormes esfuerzos de personas extraordinariamente inteligentes. Por ello, cuando el matemático británico Andrew Wiles resolvió al fin el problema en 1995, la magnitud de su hazaña era obvia para cualquiera. Ni siquiera hacía falta saber cuál era el problema, y mucho menos cómo lo había resuelto. Era el equivalente matemático a la primera ascensión al Everest.

Además de su importancia para las matemáticas, la solución de Wiles también envolvía una historia de gran interés humano. Cuando tenía diez años, Wiles había quedado tan intrigado por el problema que decidió hacerse matemático y resolverlo. Llevó a cabo la primera parte del plan, y llegó tan lejos como especializarse en teoría de números, el área general a la que pertenece el último teorema de Fermat. Pero cuanto más aprendía sobre matemáticas, más imposible parecía la empresa. El último teorema de Fermat era una curiosidad desconcertante, una cuestión aislada del tipo que a cualquier teórico de números se le podría ocurrir sin la más mínima evidencia convincente. No encajaba en ningún corpus de técnica

poderoso. En una carta a Heinrich Olbers, el gran Gauss lo había menospreciado, diciendo que el problema tenía «poco interés para mí, puesto que es fácil formular un montón de proposiciones semejantes, que no se pueden demostrar ni refutar»[3]. Wiles decidió que su sueño de la infancia había sido poco realista y puso a Fermat en el trastero. Pero entonces, milagrosamente, otros matemáticos hicieron un avance fundamental que vinculaba el problema con un tema nuclear en la teoría de números en el que Wiles ya era un experto. Gauss, de forma inhabitual en él, había subestimado la importancia del problema y no había sido consciente de que podía vincularse con una profunda, aunque aparentemente no relacionada, área de las matemáticas.

Establecido este vínculo, Wiles podía ahora trabajar en el enigma de Fermat y hacer al mismo tiempo investigación respetable en la moderna teoría de números. Mejor aún, si Fermat no lo resolvió, cualquier cosa importante que él descubriera mientras trataba de demostrarlo sería publicable por sí misma. De modo que Fermat salió del trastero y Wiles empezó a pensar en serio en el problema. Tras siete años de investigación obsesiva, realizada en privado y en secreto —una precaución inusual en matemáticas— llegó a estar convencido de que había encontrado una solución. Presentó una serie de seminarios en una prestigiosa conferencia sobre teoría de números, bajo un oscuro título que no engañaba a nadie[4]. Así estalló la excitante noticia, tanto en los medios de comunicación

[3] Gauss, carta a Heinrich Olbers, 21 de marzo de 1816.
[4] El título de Wiles era «Curvas modulares, formas elípticas y representaciones de Galois».

como en los círculos académicos: el último teorema de Fermat había sido demostrado.

La demostración era impresionante y elegante, estaba llena de buenas ideas. Por desgracia, los expertos descubrieron pronto una seria laguna en su argumentación. Esta circunstancia es tristemente común en los intentos de demoler grandes problemas no resueltos de las matemáticas, y casi siempre se prueba fatal. Sin embargo, por una vez los hados fueron amables. Con la ayuda de su antiguo estudiante Richard Taylor, Wiles consiguió salvar la laguna, corregir la demostración y completar su solución. La carga emocional implicada se hizo clara de forma muy gráfica en el programa de televisión: debe haber sido la única ocasión en que un matemático ha roto en lágrimas en pantalla, con solo recordar los traumáticos sucesos y el triunfo final.

Quizá haya usted advertido que yo no le he contado cuál *es* el último teorema de Fermat. Lo he hecho adrede; el teorema se tratará en el lugar apropiado. Por lo que respecta al programa de televisión, no importa realmente. De hecho, a los matemáticos nunca les ha preocupado gran cosa si el teorema que Fermat garabateó en su margen es verdadero o falso, porque no hay nada de gran relevancia que dependa de la respuesta. Así que ¿por qué todo ese revuelo? Porque mucho depende de la incapacidad de la comunidad matemática para *encontrar* la respuesta. No es solo un golpe a nuestra autoestima: significa que a las teorías matemáticas existentes les falta algo vital. Además, el teorema es muy fácil de

enunciar y esto se suma a su aire de misterio. ¿Cómo algo que parece tan simple resulta ser tan difícil?

Aunque a los matemáticos no les preocupara en realidad la respuesta, sí les preocupaba profundamente no saber cuál era. Y aún más les preocupaba encontrar un método que pudiera darla, porque ello arrojaría luz no solo sobre la pregunta de Fermat, sino sobre muchas otras. Esto suele suceder con los grandes problemas matemáticos: son los métodos utilizados para resolverlos, antes que los propios resultados, lo que más cuenta. Por supuesto, a veces también importa el resultado real: depende de cuáles sean sus consecuencias.

La solución de Weil es demasiado complicada y técnica para la televisión; de hecho, los detalles solo son accesibles para los especialistas[5]. La demostración implica una bonita historia matemática, como veremos en su momento, pero cualquier intento por explicarla en televisión habría perdido de inmediato a la mayoría de la audiencia. En su lugar, el programa se concentró de modo razonable en una pregunta más personal: ¿cómo se aborda un problema matemático notoriamente difícil que lleva un enorme bagaje histórico? A los espectadores se les mostraba que, diseminado a lo largo del globo, existía un grupo, pequeño pero entregado, de matemáticos con una profunda preocupación por su área de investigación, que se comunicaban, tomaban nota del trabajo de los demás y dedicaban gran parte de su vida a hacer avanzar el conocimiento matemático. Su inversión emocional y su

[5] Andrew Wiles, «Modular elliptic curves and Fermat's last theorem», *Annals of Mathematics*, n.º 141 (1995), pp. 443-551.

interacción social quedaban claramente de manifiesto. No eran autómatas inteligentes, sino personas reales, comprometidas con su tema de estudio. Ese era el mensaje.

Esas son tres grandes razones por las que el programa tuvo tanto éxito: un problema importante, un héroe con una maravillosa historia humana y unos actores secundarios implicados emocionalmente. Pero sospecho que había una cuarta razón, no tan encomiable. La mayoría de los ajenos a las matemáticas apenas oyen hablar de nuevos desarrollos en la disciplina, y ello por una variedad de razones más que comprensibles: no están tremendamente interesados; los periódicos apenas mencionan algo que tenga que ver con las matemáticas; cuando lo hacen, suele ser jocoso o trivial; y no muchas cosas en la vida diaria parecen estar afectadas por lo que estén haciendo los matemáticos entre bastidores. Demasiado a menudo las matemáticas escolares se presentan como un libro cerrado en donde toda pregunta tiene una respuesta. Los estudiantes fácilmente pueden llegar a imaginar que matemáticas nuevas son tan raras como dientes de gallina.

Desde este punto de vista, la gran noticia no era que se hubiera demostrado el último teorema de Fermat. Era que al final *alguien había hecho unas matemáticas nuevas*. Puesto que los matemáticos han necesitado más de trescientos años para encontrar una solución, muchos espectadores concluyeron subconscientemente que este gran avance era las primeras nuevas matemáticas descubiertas en los últimos trescientos años. No estoy sugiriendo que lo creyeran *explícitamente*. Deja de ser una posición sostenible

en cuanto se hacen algunas preguntas obvias, tales como «¿por qué el gobierno gasta dinero en los departamentos de matemáticas en la universidad?». Pero subconscientemente era una hipótesis común por defecto, no cuestionada ni examinada. Hacía aún mayor la magnitud del logro de Wiles.

Uno de los objetivos de este libro es mostrar que la investigación matemática es floreciente y que de continuo se hacen nuevos descubrimientos. No se oye hablar mucho de esta actividad porque la mayor parte es demasiado técnica para los no especialistas, porque la mayoría de los medios de comunicación recelan de cualquier cosa intelectualmente más desafiante que *Factor X*, y porque las aplicaciones de las matemáticas se ocultan de forma deliberada para no causar alarma: «¿Qué? ¿Mi iPhone depende de matemáticas avanzadas? ¿Cómo voy a entrar en Facebook si me suspendieron en matemáticas?».

La historia muestra que nuevas matemáticas suelen aparecer a partir de descubrimientos en otras áreas. Cuando Isaac Newton estableció sus leyes del movimiento y su ley de la gravedad, que juntas describen el movimiento de los planetas, no dio fin al problema de entender el Sistema Solar. Por el contrario, los matemáticos tuvieron que tratar con todo un nuevo rango de preguntas: sí, conocemos las leyes, pero ¿qué implican? Newton inventó el cálculo infinitesimal para responder a esa pregunta, pero su nuevo método también tiene limitaciones. A menudo reformula la pregunta en lugar de ofrecer la respuesta. Convierte el problema en una expresión de un tipo especial, llamada ecuación diferencial,

cuya *solución* es la respuesta. Pero aún hay que resolver la ecuación. No obstante, el cálculo infinitesimal era un comienzo brillante. Nos mostraba que las respuestas eran posibles, y proporcionaba una forma efectiva para buscarlas, que sigue proporcionando ideas importantes más de trescientos años después.

A medida que crecía el conocimiento matemático de la humanidad, una segunda fuente de inspiración empezó a desempeñar un papel cada vez mayor en la creación de nuevo conocimiento: las demandas internas de las propias matemáticas. Si, por ejemplo, uno sabe cómo resolver ecuaciones algebraicas de primero, segundo, tercero y cuarto grado, entonces no se necesita mucha imaginación para preguntarse por el quinto grado. (El grado es básicamente una medida de complejidad, pero ni siquiera se necesita saber lo que es para plantear la pregunta obvia). Si una solución se muestra evasiva, como sucedía en este caso, ese hecho *por sí solo* hace a los matemáticos aún más determinados a encontrar una respuesta, independientemente de que el resultado tenga o no aplicaciones útiles.

No estoy sugiriendo que las aplicaciones no importan. Pero si una parte muy especial de las matemáticas sigue apareciendo en preguntas sobre la física de las ondas —ondas oceánicas, vibraciones, sonido, luz—, entonces sin duda tiene sentido investigar esa herramienta por sí misma. No hace falta saber por adelantado cómo se va a utilizar exactamente una nueva idea: las ondas son comunes a tantas áreas de interés que las nuevas ideas están abocadas a ser útiles para algo. En este caso, ese algo incluía

a la radio, la televisión y el radar[6]. Si alguien piensa en una nueva forma de entender el flujo de calor, y da con una nueva y brillante técnica que lamentablemente carece de apoyo matemático adecuado, entonces tiene sentido clasificar el conjunto *como una parte de las matemáticas*. Incluso si a uno le importa un bledo cómo fluye el calor, el resultado podría ser perfectamente aplicable en otro lugar. El análisis de Fourier, que surgió de esta línea de investigación concreta, es presumiblemente la idea matemática más útil que se haya encontrado. Está en la base de las modernas telecomunicaciones, hace posibles las cámaras digitales, ayuda a limpiar viejas películas y grabaciones, y una extensión moderna es utilizada por el FBI para almacenar registros de huellas dactilares[7].

Después de miles de años de este tipo de intercambio entre los usos externos de las matemáticas y su estructura interna, estos dos aspectos de la disciplina han llegado a estar tan densamente entretejidos que separarlos es casi imposible. No obstante, las actitudes mentales implicadas son más fácilmente distinguibles, lo que lleva a una clasificación general de las matemáticas en dos tipos: puras y aplicadas. Esto se puede defender como una tosca manera de situar las ideas matemáticas en el paisaje intelectual, pero no es una descripción muy precisa de la propia disciplina. En el mejor de los casos distingue dos extremos de un espectro continuo de estilos matemáticos. En el peor, da una imagen falsa de qué partes de la disciplina son útiles y de dónde proceden las ideas. Como sucede con todas las ramas de la ciencia, lo que da a las

[6] Ian Stewart, *Seventeen Equations that Changed the World*, Profile, 2012, capítulo 11.
[7] *Ibid*, capítulo 9.

matemáticas su poder es la *combinación* de razonamiento abstracto e inspiración procedente del mundo exterior, que se realimentan de modo mutuo. Separar los dos aspectos no solo es imposible sino que es absurdo.

La mayoría de los problemas matemáticos realmente importantes, los grandes problemas de los que trata este libro, han surgido dentro de la disciplina gracias a un tipo de ensimismamiento intelectual. La razón es sencilla: son problemas *matemáticos*. Las matemáticas se presentan a menudo como un conjunto de áreas aisladas, cada una de ellas con sus propias técnicas especiales: álgebra, geometría, trigonometría, análisis, combinatoria, probabilidad. Suele enseñarse así, por una buena razón: situar cada tema separado en una única área bien definida ayuda a los estudiantes a organizar la materia en su mente. Es una razonable primera aproximación a la estructura de las matemáticas, especialmente las matemáticas ya bien establecidas. Sin embargo, esta rígida delimitación deja de ser válida en las fronteras de la investigación. No es solo que se difuminen las fronteras entre las áreas principales de las matemáticas. Es que realmente no existen.

Todo matemático investigador es consciente de que, en cualquier momento, de forma súbita e impredecible, puede resultar que el problema en el que está trabajando requiera ideas procedentes de un área aparentemente no relacionada. De hecho, a menudo la nueva investigación combina áreas. Por ejemplo, mi propia investigación se centra sobre todo en formación de patrones en sistemas dinámicos, sistemas que cambian con el tiempo de

acuerdo con reglas específicas. Un ejemplo típico es la forma en que se mueven los animales. Un caballo al trote repite la misma secuencia de movimientos de las patas una y otra vez, y hay una pauta clara: las patas que están relacionadas en diagonal tocan el suelo al mismo tiempo. Es decir, primero tocan las patas delantera izquierda y trasera derecha, y luego las otras dos. ¿Es este un problema sobre pautas, en cuyo caso los métodos adecuados proceden de la teoría de grupos, el álgebra de la simetría? ¿O es un problema de dinámica, en cuyo caso el área apropiada son las ecuaciones diferenciales al estilo newtoniano?

La respuesta es que, por definición, tiene que ser ambas cosas. No es su intersección, que sería el material que tienen en común —básicamente, nada—. En su lugar es una nueva «área», a caballo entre dos de las tradicionales divisiones de las matemáticas. Es como un puente a través de un río que separa dos países: une los dos, pero no pertenece a ninguno de ellos. Pero este puente no es una estrecha franja de carretera; su tamaño es comprable al de cada uno de los países. Y lo que es aún más esencial, los métodos implicados no se limitan a los de las dos áreas. De hecho, prácticamente todos los cursos de matemáticas que yo he estudiado han tenido un papel en alguna parte de mi investigación. La asignatura que cursé sobre teoría de Galois cuando yo era estudiante en Cambridge trataba de cómo resolver (más exactamente, por qué no podemos resolver) una ecuación algebraica de quinto grado. El curso de teoría de grafos trataba de redes, puntos unidos por líneas. Nunca hice un curso de sistemas

dinámicos, porque mi doctorado era en álgebra, pero con los años reuní las ideas básicas, desde los estados estacionarios al caos. Teoría de Galois, teoría de grafos, sistemas dinámicos: tres áreas independientes. O así lo suponía hasta 2011, cuando trataba de entender cómo detectar dinámica caótica en una red de sistemas dinámicos y un paso crucial dependía de cosas que había aprendido cuarenta y cinco años antes en mi curso sobre teoría de Galois.

Así pues, las matemáticas no son como un mapa político del mundo, con cada especialidad claramente rodeada de una frontera nítida, cada país claramente distinguido de sus vecinos por estar coloreado de rosa, verde o azul claro. Son más parecidas a un paisaje natural, donde nunca se puede decir con precisión dónde termina el valle y empieza la montaña, dónde el bosque se funde con la maleza y llanuras con hierba, dónde los lagos insertan regiones de agua en otros tipos de terreno, dónde los ríos unen las pendientes nevadas de las montañas con los distantes océanos. Pero este paisaje matemático en cambio continuo no consiste en rocas, agua y plantas sino en ideas; está ligado no por la geografía sino por la lógica. Y es un paisaje dinámico, que varía a medida que se descubren o inventan nuevas ideas o métodos. Los conceptos importantes con amplias implicaciones son como cimas montañosas; las técnicas con muchos usos son como ríos caudalosos que llevan a los viajeros a través de llanuras fértiles. Cuanto más claramente definido está el paisaje, más fácil es detectar picos vírgenes o terreno inexplorado que crea obstáculos

indeseados. Con el tiempo, algunos de los picos y obstáculos adquieren un estatus de icono. Estos son los grandes problemas.

¿Qué es lo que hace grande a un gran problema matemático? Es la profundidad intelectual, combinada con simplicidad y elegancia. Más aún: tiene que ser *difícil*. Cualquiera puede escalar una loma; el Everest es otra cosa muy diferente. Un gran problema es normalmente fácil de enunciar, aunque los términos requeridos pueden ser elementales o pueden ser muy técnicos. Los enunciados del último teorema de Fermat y del problema de los cuatro colores tienen sentido inmediato para cualquiera que esté familiarizado con las matemáticas de la escuela. Por el contrario, es imposible enunciar siquiera la conjetura de Hodge o la hipótesis del hueco de masas sin invocar conceptos profundos en las fronteras de la investigación —después de todo, la última procede de la teoría cuántica de campos—. Sin embargo, para los versados en tales áreas, el enunciado de la cuestión concernida es simple y natural, no implica páginas y páginas de texto denso e impenetrable. En medio hay problemas que requieren algo al nivel de matemáticas universitarias si se quieren entender con todo detalle. Una idea más general de la esencia del problema —de dónde procede, por qué es importante, qué se podría hacer si se dispusiera de una solución— es normalmente accesible para cualquier persona interesada, y eso es lo que estoy tratando de ofrecer. Admito que la conjetura de Hodge es un hueso difícil de roer a este respecto, porque es muy técnica y muy abstracta. Sin embargo, es uno de los siete problemas

matemáticos del milenio del Instituto Clay, con un premio asociado de un millón de dólares, y necesariamente debe ser incluido.

Los grandes problemas son creativos: ayudan al nacimiento de nuevas matemáticas. En 1900 David Hilbert pronunció una conferencia en el Congreso Internacional de Matemáticos en París, en la que hizo una lista de 23 de los problemas más importantes en matemáticas. No incluía el último teorema de Fermat pero lo mencionaba en su introducción. Cuando un matemático distinguido hace una lista de los que él piensa que son algunos de los grandes problemas, otros matemáticos prestan atención. Los problemas no estarían en la lista a menos que fueran importantes y difíciles. Es natural plantearse el desafío y tratar de responderlos. Desde entonces, resolver uno de los problemas de Hilbert ha sido una buena manera de ganarse los galones matemáticos. Muchos de estos problemas son demasiado técnicos para incluirlos aquí, muchos son programas abiertos antes que problemas específicos, y varios aparecen más tarde por derecho propio. Pero merecen ser mencionados, por lo que he puesto un breve resumen en las notas.[8]

[8] Los problemas de Hilbert, y su estatus actual, tal como figuran con ligeros cambios en *Profesor Stewart's Hoard of Mathematical Treasures*, Profile, 2009 (Hay traducción en castellano: *Baúl de tesoros matemáticos*, Crítica, Barcelona, 2010.), son los siguientes:

1. **Hipótesis del continuo**: ¿Hay un número cardinal infinito estrictamente comprendido entre las cardinalidades de los enteros y de los números reales? Resuelto por Paul Cohen en 1963: la respuesta depende de los axiomas que se utilicen para la teoría de conjuntos.
2. **Consistencia lógica de la aritmética**: Demostrar que los axiomas estándar de la aritmética no pueden llevar a una contradicción. Resuelto por Kurt Gödel en 1931: imposible con los axiomas habituales para la teoría de conjuntos.
3. **Igualdad de volúmenes de tetraedros**: Si dos tetraedros tienen el mismo volumen, ¿es siempre posible dividir uno de ellos en un número finito de fragmentos poligonales y reensamblarlos para formar el otro? Resuelto en 1901 por Max Dehn, en sentido negativo.
4. **La línea recta como la distancia más corta entre dos puntos**: Formular axiomas para la geometría en términos de la definición anterior de «línea recta» e investigar las implicaciones. Demasiado general para tener una solución definitiva, pero se ha trabajado mucho en ello.

5. **Grupos de Lie sin suponer diferenciabilidad**: Cuestión técnica en la teoría de grupos de transformaciones. En una interpretación, resuelto por Andrew Gleason en los años cincuenta del siglo pasado. En otra interpretación, resuelto por Hidehiko Yamabe.
6. **Axiomas para la física**: Desarrollar un sistema de axiomas riguroso para las áreas matemáticas de la física, tales como la probabilidad y la mecánica. Andrei Kolmogorov axiomatizó la probabilidad en 1933.
7. **Números irracionales y trascendentes**: Demostrar que ciertos números son irracionales o trascendentes. Resuelto por Aleksandr Gelfond y Theodor Schneider en 1934.
8. **Hipótesis de Riemann**: Demostrar que todos los ceros no triviales de la función zeta de Riemann yacen en la recta crítica. Ver capítulo 9.
9. **Leyes de reciprocidad en campos de números**: Generalizar la ley clásica de reciprocidad cuadrática, sobre cuadrados respecto a un módulo, a potencias más altas. Parcialmente resuelto.
10. **Determinar cuándo una ecuación diofántica tiene solución**: Encontrar un algoritmo que, cuando se aplica a una ecuación polinómica con varias variables, determina si existe una solución en números enteros. Demostrado imposible por Yuri Matiyasevich en 1970.
11. **Formas cuadráticas con números algebraicos como coeficientes**: Cuestiones técnicas sobre la solución de ecuaciones diofánticas con varias variables. Parcialmente resuelto.
12. **Teorema de Kronecker sobre campos abelianos**: Cuestiones técnicas que generalizan un teorema de Kronecker. Aún no resuelto.
13. **Resolver ecuaciones de séptimo grado utilizando funciones especiales**: Demostrar que la ecuación general de séptimo grado no puede resolverse utilizando funciones de dos variables. Una interpretación refutada por Andrei Kolmogorov y Vladimir Arnold.
14. **Finitud de sistemas completos de funciones**: Extender un teorema de Hilbert sobre invariantes algebraicos a todos los grupos de transformaciones. Demostrado falso por Masayoshi Nagata en 1959.
15. **Cálculo enumerativo de Schubert**: Hermann Schubert encontró un método no riguroso para contar varias configuraciones geométricas. Hacer el método riguroso. Todavía no hay una solución completa.
16. **Topología de curvas y superficies**: ¿Cuántas componentes conexas puede tener una curva algebraica de un grado dado? ¿Cuántos ciclos periódicos distintos puede tener una ecuación diferencial algebraica de un grado dado? Progresos limitados.
17. **Expresar formas definidas por cuadrados**: Si una función racional toma siempre valores no negativos, ¿debe ser una suma de cuadrados? Resuelto por Emil Artin, D. W. Dubois y Albrecht Pfister. Verdadero sobre los números reales, falso en algunos otros sistemas de números.
18. **Teselar el espacio con poliedros**: Cuestiones generales sobre llenar el espacio con poliedros congruentes. También menciona la conjetura de Kepler, ahora demostrada; véase el capítulo 5.
19. **Analiticidad de soluciones en el cálculo de variaciones**: El cálculo de variaciones responde a preguntas como: «Encontrar la curva más corta con las siguientes propiedades». Si un problema semejante está definido por funciones simples, ¿debe ser también simple la solución? Demostrado por Ennio de Giorgi en 1957, y por John Nash.
20. **Problemas de valores de contorno**: Entender las soluciones de las ecuaciones diferenciales de la física, dentro de una región del espacio, cuando están prescritas las propiedades de la solución en la frontera de dicha región. Esencialmente resuelto, por numerosos matemáticos.
21. **Existencia de ecuaciones diferenciales con monodromía dada**: Un tipo especial de ecuación diferencial compleja puede entenderse en términos de sus puntos singulares y su grupo de monodromía. Demostrar que cualquier combinación de estos datos puede ocurrir. Respondida sí o no, dependiendo de la interpretación.
22. **Uniformización utilizando funciones automorfas**: Cuestión técnica sobre simplificación de ecuaciones. Resuelto por Paul Koebe poco después de 1900.

Eso es lo que hace grande a un gran problema matemático. Raras veces lo que lo hace problemático es decidir cuál debería ser la respuesta. Para prácticamente todos los grandes problemas, los matemáticos tienen una idea muy clara de cuál debería ser la respuesta —o la tenían, si ahora se conoce una solución—. De hecho, el enunciado del problema incluye a veces la respuesta esperada. Eso es lo que sucede con una conjetura: una hipótesis plausible, basada en una amplia evidencia. Muchas conjeturas bien estudiadas resultan ser finalmente correctas, aunque no todas. Términos más antiguos como hipótesis tienen el mismo significado, y en el caso de Fermat la palabra «teorema» es (más exactamente, era) un abuso de lenguaje: un teorema requiere una demostración, pero eso era precisamente lo que faltaba hasta que llegó Wiles.

La demostración es, de hecho, el requisito que hace problemáticos los grandes problemas. Cualquiera moderadamente competente puede realizar algunos cálculos, detectar una pauta aparente y destilar su esencia en un conciso enunciado. Los matemáticos exigen más pruebas que eso: insisten en una demostración completa, lógicamente impecable. O, si la respuesta resulta ser negativa, una refutación. En realidad, no es posible apreciar el encanto seductor de un gran problema sin apreciar el papel vital de la demostración en la empresa matemática. Cualquiera puede hacer una conjetura razonada. Lo difícil es demostrar que es correcta. O que es falsa.

---

23. **Desarrollo del cálculo de variaciones**: Hilbert pedía nuevas ideas en el cálculo de variaciones. Se ha hecho mucho trabajo; cuestión demasiado vaga para considerarse resuelta.

El concepto de demostración matemática ha cambiado en el curso de la historia, y los requisitos lógicos se han hecho generalmente más estrictos. Ha habido muchas sesudas discusiones filosóficas sobre la naturaleza de la demostración que han planteado algunas cuestiones importantes. Se han propuesto y puesto en práctica definiciones lógicas precisas de «demostración». La que enseñamos a los estudiantes en la universidad es que una demostración empieza con una colección de hipótesis explícitas llamadas axiomas. Los axiomas son, por así decir, las reglas del juego. Otros axiomas son posibles, pero llevan a juegos diferentes. Fue Euclides, el antiguo geómetra griego, quien introdujo este enfoque de las matemáticas, que sigue siendo válido hoy. Acordados los axiomas, una demostración de un enunciado es una serie de pasos, cada uno de los cuales es una consecuencia lógica de alguno de los axiomas, o de enunciados previamente demostrados, o de ambos. En efecto, el matemático está explorando un laberinto lógico cuyos cruces son enunciados y cuyos pasillos son deducciones válidas. Una demostración es un camino a través del laberinto, partiendo de los axiomas. Lo que demuestra es el enunciado en el que acaba.

No obstante, este rígido concepto de demostración no es toda la historia. Ni siquiera es la parte más importante de la historia. Es como decir que una sinfonía es una secuencia de tonos musicales, sujetos a las reglas de la armonía. Carece de creatividad. No nos dice cómo encontrar demostraciones, ni siquiera cómo validar las demostraciones de otras personas. No nos dice qué lugares del laberinto son importantes. No nos dice qué caminos son elegantes y

cuáles son feos, cuáles son importantes y cuáles irrelevantes. Es una descripción formal y mecánica de un proceso que tiene muchos otros aspectos, en especial una dimensión humana. Las demostraciones son descubiertas por personas, y la investigación en matemáticas no es solo una cuestión de lógica paso a paso.

Tomar al pie de la letra la definición formal de demostración puede llevar a demostraciones que son prácticamente ilegibles, porque la mayor parte del tiempo se pasa poniendo puntos lógicos sobre las íes y cruces lógicos en las tes en circunstancias en las que el resultado ya se tiene delante. Por ello los matemáticos en activo toman atajos y dejan fuera lo que es rutinario u obvio. Dejan claro que hay una laguna utilizando frases tópicas como «es fácil verificar que» o «cálculos rutinarios implican». Lo que no hacen, al menos no de modo consciente, es introducir una dificultad lógica y simular que no existe. De hecho, un matemático competente se apartará de su camino para señalar con exactitud aquellas partes del argumento que son lógicamente frágiles, y dedicará la mayor parte de su tiempo a explicar cómo hacerlas suficientemente robustas. El resultado es que una demostración, en la práctica, es una historia matemática con su propio flujo narrativo. Tiene un comienzo, un desarrollo y un final. A menudo tiene subargumentos, que nacen del argumento principal, cada uno con su propia resolución. El matemático británico Christopher Zeeman comentó en una ocasión que un teorema es un punto de reposo intelectual. Uno puede detenerse, recuperar aliento y sentir que ha llegado a un lugar definido. Los subargumentos atan un cabo suelto en la historia

principal. Las demostraciones se parecen a narraciones en otros aspectos: a menudo tienen uno o más protagonistas —ideas en lugar de personas, por supuesto— cuyas interacciones complejas llevan a la revelación final.

Como indica la definición que se da a los estudiantes, una demostración parte de una hipótesis claramente enunciada, deduce consecuencias lógicas de una manera coherente y estructurada, y termina en lo que sea que se quiere demostrar. Pero una demostración no es solamente una lista de deducciones, y la lógica no es el único criterio. Una demostración es una historia contada y diseccionada por personas que han pasado buena parte de su vida aprendiendo a leer historias y a encontrar errores o inconsistencias: personas cuyo principal objetivo es demostrar que el narrador está *equivocado*, y que poseen el asombroso don de detectar debilidades y corregirlas hasta que se deshacen en una nube de polvo. Si algún matemático pretende haber resuelto un problema importante, ya sea uno grande o algo digno aunque menos excelso, el reflejo profesional no es gritar ¡hurra! y descorchar una botella de champán, sino tratar de rebatirlo.

Esto puede sonar negativo, pero la demostración es la única herramienta fiable que tienen los matemáticos para asegurar que lo que dicen es correcto. Anticipando este tipo de respuesta, los investigadores dedican muchos esfuerzos a tratar de rebatir sus propias ideas y demostraciones. Así es menos embarazoso. Cuando la historia ha sobrevivido a este tipo de escrutinio crítico, el consenso pronto se convierte en un acuerdo de que es correcto, y en

ese momento el inventor de la demostración recibe la alabanza, crédito y recompensa adecuados. En cualquier caso, eso es lo que sucede normalmente, aunque no siempre puede parecerlo así para las personas implicadas. Si uno está próximo a la acción, su imagen de lo que está sucediendo puede ser diferente de la de un observador más distanciado.

* * * *

¿Cómo resuelven problemas los matemáticos? Ha habido pocos estudios científicos rigurosos sobre esta cuestión. La moderna investigación en educación, basada en la ciencia cognitiva, se centra sobre todo en la educación hasta el nivel del instituto. Algunos estudios abordan la enseñanza de las matemáticas en la universidad, pero son relativamente escasos. Hay diferencias importantes entre aprender y enseñar las matemáticas ya existentes y crear nuevas matemáticas. Muchos de nosotros podemos tocar un instrumento musical, pero muchos menos podemos componer un concierto o siquiera escribir una canción pop.

Cuando se llega a la creatividad en los niveles más altos, mucho de lo que sabemos —o pensamos que sabemos— procede de la introspección. Pedimos a los matemáticos que expliquen sus procesos mentales y buscamos principios generales. Uno de los primeros intentos serios de descubrir cómo piensan los matemáticos fue *The Psychology of Invention in the Mathematical Field* de Jacques Hadamard, publicado por primera vez en 1945[9]. Hadamard entrevistó a destacados matemáticos y científicos de su tiempo y les

[9] Reimpreso como: Jacques Hadamard, *The Psychology of Invention in the Mathematical Field*, Dover, 1954.

pidió que describieran cómo pensaban cuando trabajaban en problemas difíciles. Lo que emergió, con mucha fuerza, era el papel vital de lo que a falta de un término mejor debe describirse como intuición. Alguna propiedad de la mente subconsciente guiaba sus pensamientos. Sus ideas más creativas no surgieron mediante una lógica paso a paso, sino por repentinos saltos incontrolados.

Una de las descripciones más detalladas de este enfoque aparentemente ilógico de las cuestiones lógicas la proporcionó el matemático francés Henri Poincaré, una de las figuras destacadas de finales del siglo XIX y principios del siglo XX. Poincaré recorrió buena parte de las matemáticas fundando varias áreas nuevas y cambiando radicalmente muchas otras. Desempeña un papel destacado en varios capítulos posteriores. También escribió libros de divulgación científica, y esta amplia experiencia puede haberle ayudado a obtener una comprensión más profunda de sus propios procesos mentales. En cualquier caso, Poincaré insistía en que la lógica consciente era solo una parte del proceso creativo. Sí, había momentos en que era indispensable decidir cuál era en realidad el problema, verificar de modo sistemático la respuesta. Pero en medio, Poincaré sentía que su cerebro estaba trabajando a menudo en el problema sin decírselo, de maneras que él sencillamente no podía imaginar.

Su esbozo del proceso creativo distinguía tres etapas: preparación, incubación e iluminación. La preparación consiste en esfuerzos lógicos conscientes para fijar el problema, hacerlo preciso y atacarlo por métodos convencionales. Poincaré consideraba esencial esta

fase: pone en marcha el subconsciente y le proporciona materia prima para que trabaje con ella. La incubación tiene lugar cuando uno deja de pensar en el problema y hace alguna otra cosa. Ahora el subconsciente empieza a combinar ideas, a menudo ideas descabelladas, hasta que empieza a verse la luz. Con suerte, esto lleva a la iluminación: su subconsciente le da palmadas en el hombro y en su cerebro se enciende la proverbial bombilla.

Este tipo de creatividad es como andar sobre una cuerda tensa. Por una parte, uno no quiere resolver un problema difícil a menos que esté familiarizado con el área a la que parece pertenecer —y con muchas otras áreas, que pueden estar relacionadas o no, por si acaso lo están—. Por otra parte, si todo lo que uno hace es quedarse atrapado en modos de pensamiento estándar, que otros ya han ensayado infructuosamente, entonces quedará atascado en una rutina mental y no descubrirá nada nuevo. De modo que el truco es saber muchas cosas, integrarlas de modo consciente, poner en marcha el cerebro durante semanas... y luego dejar de lado la cuestión. La parte intuitiva de su mente empieza a trabajar, frota unas ideas con otras para ver si salta la chispa y le notifica cuando ha encontrado algo. Esto puede suceder en cualquier momento: Poincaré vio de pronto cómo resolver un problema que le había estado preocupando durante meses al subir a un autobús. A Srinivasa Ramanujan, un matemático indio autodidacta con talento para fórmulas extraordinarias, las ideas le llegaban a veces en sueños. Arquímedes es famoso por haber descubierto la forma de probar si un metal era oro cuando se estaba bañando.

Poincaré puso especial cuidado en señalar que sin el período inicial de preparación, el progreso es poco probable. El subconsciente, insistía, necesita que se le dé mucho en qué pensar, o de lo contrario no se pueden formar las fortuitas combinaciones de ideas que finalmente lleven a una solución. La preparación genera inspiración. Él también debe haber sabido —puesto que cualquier matemático creativo lo sabe— que este sencillo proceso de tres etapas difícilmente ocurre solo una vez. Resolver un problema requiere a manudo más de un avance fundamental. La etapa de incubación para una idea puede ser interrumpida por un proceso subsidiario de preparación, incubación e iluminación de algo que es necesario para que funcione la primera idea. La solución a cualquier problema que se precie de tal, sea grande o no, implica por lo general muchas de tales secuencias, anidadas unas dentro de otras como uno de los intrincados fractales de Benoit Mandelbrot. Uno resuelve un problema dividiéndolo en subproblemas. Uno se convence de que si puede resolver estos subproblemas, podrá reunir los resultados para resolver el conjunto. Entonces elabora los subproblemas. A veces resuelve uno; a veces fracasa y es necesario un replanteamiento. A veces un subproblema se divide en más piezas. Tan solo seguir la pista del plan puede ser una gran tarea.

He descrito el funcionamiento del subconsciente como «intuición». Esta es una de esas palabras seductoras como «instinto», que es ampliamente utilizada incluso si carece de significado real. Es un nombre para algo cuya presencia reconocemos, pero que no entendemos. La intuición matemática es la capacidad de la mente

para sentir forma y estructura, para detectar pautas que no podemos percibir de forma consciente. La intuición carece de la claridad cristalina de la lógica consciente, pero lo compensa con dirigir la atención a cosas que nunca habríamos considerado de forma consciente. Los neurocientíficos apenas están empezando a entender cómo el cerebro realiza tareas mucho más simples. Pero como quiera que trabaje la intuición, debe ser una consecuencia de la estructura del cerebro y de cómo interacciona con el mundo exterior.

A menudo la contribución clave de la intuición es hacernos conscientes de los puntos débiles de un problema, lugares en donde puede ser vulnerable a un ataque. Una demostración matemática es como una batalla, o si usted prefiere una metáfora menos bélica, como un juego de ajedrez. Una vez que se ha identificado un punto débil, puede explotarse gracias al dominio técnico que tiene el matemático de la maquinaria de las matemáticas. Como Arquímedes, que quería un punto de apoyo firme para poder mover la Tierra, el matemático investigador necesita alguna manera de ejercer palanca sobre el problema. Una idea clave puede abrirlo y hacerlo vulnerable a métodos estándar. Después de eso, es solo cuestión de técnica.

Mi ejemplo favorito de este tipo de palanca es un rompecabezas que no tiene una importancia matemática intrínseca, pero lleva un mensaje importante. Suponga que usted tiene un tablero de ajedrez, con 64 casillas, y un suministro de fichas de dominó del tamaño preciso para cubrir dos casillas adyacentes del tablero. Entonces es

fácil cubrir todo el tablero con 32 fichas. Pero supongamos ahora que se han eliminado dos esquinas diagonalmente opuestas del tablero, como en la Figura 1. ¿Pueden cubrirse las 62 casillas restantes utilizando 31 fichas? Si uno experimenta, nada parece funcionar. Por otra parte, es difícil ver una razón obvia para que la tarea sea imposible. Hasta que uno se da cuenta de que, cualquiera que sea la forma en que se dispongan las fichas, cada una de ellas debe cubrir una casilla negra y una casilla blanca. Esta es su palanca; todo lo que tiene que hacer ahora es manejarla. Implica que cada región cubierta por fichas contiene el mismo número de casillas negras que de casillas blancas. Pero las casillas diagonalmente opuestas tienen el mismo color, de modo que eliminar dos de ellas (aquí blancas) lleva a una figura con dos casillas negras más que casillas blancas.

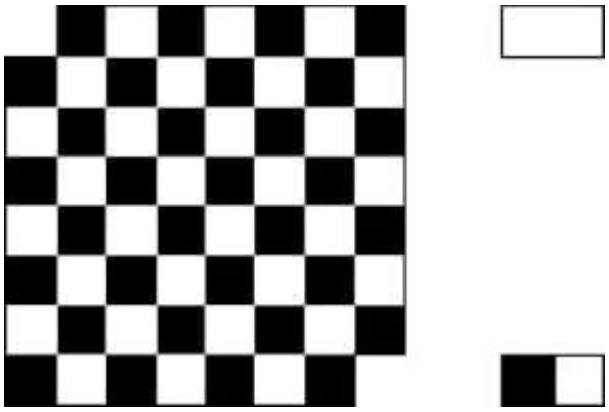


Figura 1. ¿Se puede cubrir el tablero de ajedrez recortado con fichas de dominó, cada una de las cuales cubre dos casillas (parte superior derecha)? Si se colorea la ficha de dominó (parte inferior derecha) y se cuenta cuántas casillas blancas y negras hay, la respuesta es clara.

Así que esta figura no puede cubrirse. La observación sobre la combinación de colores que cubre *cualquier* ficha es el punto débil del rompecabezas. Le da un lugar donde apoyar su palanca lógica y presionar. Si usted fuera un señor medieval asaltando un castillo, este sería el punto débil del muro, el lugar en donde debería concentrar la potencia de fuego de sus catapultas o cavar un túnel para minarlo.

La investigación en matemáticas difiere de una batalla en un aspecto importante. Cualquier territorio que uno ocupa una vez, permanece en su poder para siempre. Usted puede decidir concentrar sus esfuerzos en algún otro lugar, pero una vez que un teorema está demostrado, ya no desaparece. Así es como los matemáticos avanzan en un problema, incluso cuando no consiguen resolverlo. Ellos establecen un nuevo hecho, que queda entonces a disposición de cualquier otro y en cualquier contexto. A menudo la lanzadera para un nuevo asalto a un viejo problema surge de una joya previamente inadvertida semienterrada en un montón informe de hechos revueltos. Y esa es una razón por la que nuevas matemáticas pueden ser importantes por sí mismas, incluso si sus usos no son evidentes a primera vista. Es una parte más de territorio ocupado, un arma más en el arsenal. Su momento puede estar aún por llegar; pero ciertamente no lo hará si se estima «inútil» y se olvida, o si nunca se le permite nacer porque nadie puede ver *para qué* sirve.

Capítulo 2

Territorio primo. *La conjetura de Goldbach*

Algunos de los grandes problemas se manifiestan muy pronto en nuestra educación matemática, aunque quizá no lo advirtamos. En cuanto se nos enseña la multiplicación, tropezamos con la idea de número primo. Algunos números pueden obtenerse multiplicando entre sí dos números más pequeños; por ejemplo, 6 = 2×3. Otros, tales como 5, no pueden descomponerse de esta manera; lo mejor que podemos hacer es 5 = 1×5, que no incluye dos números *más pequeños*. De los números que pueden descomponerse se dice que son compuestos; los que no pueden son primos. Los números primos parecen cosas muy simples. En cuanto uno puede multiplicar números enteros, puede entender lo que es un número primo. Los primos son los bloques constituyentes básicos de los números naturales y aparecen por todas las matemáticas. Son también muy misteriosos, y parecen distribuirse prácticamente al azar. No hay duda: los primos son un enigma. Quizá esto es una consecuencia de su definición: no tanto lo que son como lo que no son. Por otra parte, son fundamentales para las matemáticas, de modo que no podemos simplemente levantar las manos con horror y rendirnos. Necesitamos entender los primos y desentrañar sus más íntimos secretos.

Unas pocas propiedades son obvias. Con excepción del primo más pequeño, 2, todos los primos son impares. Con la excepción de 3, la suma de sus dígitos no puede ser múltiplo de 3. Con la excepción de

5, no pueden terminar en 5. Aparte de estas reglas, y unas pocas más sutiles, uno no puede echar una mirada a un número y detectar de inmediato si es primo. Existen fórmulas para los primos, pero en su mayor parte son tramposas: no proporcionan nueva información útil sobre los primos; simplemente son maneras más ingeniosas de codificar la definición de «primo» en una fórmula. Los primos son como las personas: son individuales y no se atienen a reglas estándar.

Durante milenios, los matemáticos han mejorado poco a poco su comprensión de los números primos, y de vez en cuando se resuelve otro gran problema sobre ellos. Sin embargo, todavía quedan muchas preguntas sin resolver. Algunas son básicas y fáciles de enunciar, otras son más esotéricas. Este capítulo discute lo que sabemos y lo que no sabemos sobre estos números exasperantes pero fundamentales. Empieza estableciendo algunos de los conceptos básicos, en particular la factorización en primos: cómo expresar un número dado como un producto de primos. Incluso este proceso familiar lleva a aguas profundas en cuanto empezamos a buscar métodos auténticamente efectivos para encontrar los factores primos de un número. Una sorpresa es que parece relativamente fácil poner a prueba un número para determinar si es primo; pero si es compuesto, encontrar sus factores primos suele ser mucho más difícil.

Habiendo expuesto las ideas básicas, pasaremos al más famoso problema no resuelto sobre los primos, la conjetura de Goldbach, con 250 años de antigüedad. Los progresos recientes sobre esta

cuestión han sido espectaculares, pero todavía no decisivos. Algunos otros problemas proporcionan una breve muestra de lo que aún está por descubrir sobre esta rica pero rebelde área de las matemáticas.

Los números primos y la factorización son familiares desde la aritmética escolar, pero la mayoría de las propiedades interesantes de los primos apenas se enseñan en ese nivel, y de hecho no se demuestra nada. Hay buenas razones para ello: las demostraciones, incluso las de propiedades en apariencia obvias, son sorprendentemente difíciles. En su lugar, a los alumnos se les enseñan algunos métodos simples para trabajar con primos, y se pone el énfasis en cálculos con números relativamente pequeños. Como resultado, nuestra experiencia temprana con los primos es algo equívoca.

Los antiguos griegos conocían algunas de las propiedades básicas de los primos y sabían cómo demostrarlas. Primos y factores son el tema principal del libro VII de los *Elementos* de Euclides, el gran clásico de la geometría. Este libro en concreto contiene una presentación geométrica de la división y la multiplicación en aritmética. Los griegos preferían trabajar con longitudes de líneas, antes que con números como tales, pero es fácil reformular sus resultados en el lenguaje de los números. Euclides se ocupa en demostrar enunciados que pueden parecer obvios: por ejemplo, la proposición 16 del libro VII demuestra que cuando se multiplican dos números, el resultado es independiente del orden en que se toman. Es decir $ab = ba$, una ley básica del álgebra.

En la aritmética de la escuela, los factores primos se utilizan para encontrar el máximo común divisor (o factor común más alto) de dos números. Por ejemplo, para encontrar el máximo común divisor de 135 y 630, los factorizamos en primos:

$$135 = 3^3 \times 5 \quad 630 = 2 \times 3^2 \times 5 \times 7$$

Luego, por cada primo, tomamos la potencia mayor que aparece en ambas factorizaciones, con lo que se obtiene $3^2 \times 5$. Esta multiplicación da 45: este es el máximo común divisor. Este procedimiento da la impresión de que la factorización en primos es necesaria para encontrar máximos comunes divisores. En realidad, la relación lógica es al revés. El libro VII, proposición 2, de los *Elementos* presenta un método para encontrar el máximo común divisor de dos números enteros sin factorizarlos. Funciona restando repetidamente el número menor del mayor, y luego aplicando un proceso similar al resto resultante y el número menor, y continuando así hasta que no hay resto. Para 135 y 630, un ejemplo típico que utiliza números no muy grandes, el proceso va así. Restamos repetidamente 135 de 630.

$$630 - 135 = 495$$
$$495 - 135 = 360$$
$$360 - 135 = 225$$
$$225 - 135 = 90$$

Puesto que 90 es menor que 135, intercambiamos los dos números 90 y 135:

$$135 - 90 = 45$$

Puesto que 45 es menor que 90, intercambiamos 45 y 90:

$$90 - 45 = 45$$
$$45 - 45 = 0$$

Por lo tanto, el máximo común divisor de 135 y 630 es 45.

Este procedimiento funciona porque en cada paso reemplaza el par de números original por un par más sencillo (uno de los números es más pequeño) que tiene el mismo máximo común divisor. Por último, uno de los números divide al otro exactamente, y en esa etapa nos detenemos. El término actual para un método computacional explícito que garantiza encontrar una respuesta a un problema dado es «algoritmo». De modo que el procedimiento de Euclides se denomina ahora el algoritmo de Euclides. Es lógicamente anterior a la factorización en primos. De hecho, Euclides utiliza su algoritmo para demostrar propiedades básicas de los factores primos, y así se hace hoy en los cursos de matemáticas en la universidad.

La proposición 30 de Euclides es vital para la empresa. En términos modernos afirma que si un primo divide al producto de dos números —lo que se obtiene multiplicándolos— entonces debe

dividir a uno de ellos. La proposición 32 afirma que o bien un número es primo o bien tiene un factor primo. Juntándolas es fácil deducir que todo número es un producto de factores primos, y que esta expresión es única salvo el orden en que se escriben los factores. Por ejemplo:

$$60 = 2\times2\times3\times5 = 2\times3\times2\times5 = 5\times3\times2\times2$$

y así sucesivamente, pero la única manera de obtener 60 es reordenar la primera factorización. No hay factorización, por ejemplo, que se parezca a 60 = 7×*algo*. La existencia de la factorización procede de la proposición 32. Si el número es primo, *stop*. Si no lo es, encontramos un factor primo, dividimos para obtener un número más pequeño, y repetimos. La unicidad procede de la proposición 30. Por ejemplo, si hubiera una factorización 60 = 7×*algo*, entonces 7 debe dividir a uno de los números 2, 3 o 5, pero no lo hace.

Aquí tengo que aclarar un punto menor pero importante: el estatus excepcional del número 1. De acuerdo con la definición que hemos dado hasta ahora, 1 es claramente primo: si tratamos de descomponerlo, lo más que podemos hacer es 1 = 1×1, que no incluye números más pequeños. Sin embargo, esta interpretación causa problemas más tarde en la teoría, y por ello en el último o los dos últimos siglos los matemáticos han añadido una restricción extra. El número 1 es tan especial que no debería considerarse como primo ni como compuesto. Es otra cosa, una unidad. Una

razón para tratar 1 como un caso especial, más que como un auténtico primo, es que si llamamos primo a 1, la unicidad falla. De hecho, 1×1 = 1 ya muestra este fallo, y 1×1×1×1×1×1×1×1 = 1 se nos restriega en las narices. Podríamos modificar la unicidad para decir «único excepto 1s extra», pero eso es simplemente otra manera de admitir que 1 es especial.

Mucho más tarde, en la proposición 20 del libro IX, Euclides demuestra otro hecho clave: «Los números primos son más que cualquier multitud definida de números primos». Es decir, el número de primos es infinito. Es un teorema maravilloso con una demostración ingeniosa, pero abría una enorme caja de Pandora. Si los primos no acaban nunca, pero no parecen tener una pauta, ¿cómo podemos describir a qué se parecen?

Tenemos que enfrentarnos a esta pregunta porque no podemos ignorar los primos. Son accidentes esenciales del paisaje matemático. Son particularmente habituales, y útiles, en la teoría de números. Esta área de las matemáticas estudia las propiedades de los números naturales. Puede sonar algo elemental, pero en realidad la teoría de números es una de las más profundas y más difíciles áreas de las matemáticas. Más tarde veremos muchas pruebas de esto. En 1801, Gauss, el más destacado teórico de números de su época —presumiblemente uno de los mejores matemáticos de todos los tiempos— escribió un libro de texto avanzado sobre teoría de números, las *Disquisitiones arithmeticae* («Investigaciones en aritmética»). Entre los temas de más alto nivel, él señalaba que no deberíamos perder de vista dos cuestiones muy

básicas: «Es sabido que el problema de distinguir los números primos de los números compuestos y resolver los últimos en sus factores primos es uno de los problemas más importantes y útiles en aritmética».

En la escuela se suele enseñar concretamente una manera de encontrar los factores primos de un número: ensayar uno tras otro todos los factores posibles hasta que se encuentre algo que vale. Si no se ha encontrado un factor para cuando se ha llegado a la raíz cuadrada del número original —más preciso, al mayor número natural que es menor o igual que la raíz cuadrada—, entonces el número es primo. Si, por el contrario, se encuentra un factor, se divide por este y se repite. Es más eficaz ensayar precisamente factores primos, lo que requiere disponer de una lista de primos. Uno se detiene en la raíz cuadrada porque el factor más pequeño de cualquier número compuesto no es mayor que su raíz cuadrada. Sin embargo, este procedimiento es desesperantemente ineficaz cuando se trata con números grandes. Por ejemplo, si el número es:

1.080.813.321.843.836.712.253

entonces su factorización en primos es:

13.929.010.429×77.594.408.257

y habría que ensayar los primeros 624.401.249 primos uno tras otro para encontrar el menor de los dos factores. Por supuesto, con

un ordenador esto es relativamente fácil, pero si empezamos con un número de cien dígitos que resulta ser el producto de dos números de cincuenta dígitos, y hacemos una búsqueda sistemática a través de primos sucesivos, el universo llegará a su fin antes de que el ordenador encuentre la respuesta.

De hecho, los ordenadores actuales pueden factorizar generalmente números de cien dígitos. Mi ordenador necesita menos de un segundo para encontrar los factores primos de $10^{99} + 1$, que se parece a 1000... 001 con 98 ceros. Es un producto de 13 primos (uno de ellos aparece dos veces), de los que el menor es 7 y el mayor es:

141.122.524.877.866.182.282.233.539.317.796.144.938.305.111.168.717

Pero si yo pido al ordenador que factorice $10^{199} + 1$, con doscientos dígitos, seguirá funcionando durante tiempo y tiempo y no llegará a ninguna parte. Pese a todo, el cálculo con cien dígitos es impresionante. ¿Cuál es el secreto? Encontrar métodos más eficaces que ensayar todos los potenciales factores primos uno detrás de otro.

Ahora sabemos mucho más que Gauss sobre el primero de sus problemas (comprobar primos) y mucho menos de lo que nos gustaría sobre el segundo (la factorización). La opinión convencional es que la comprobación del carácter primo es mucho más sencilla que la factorización. Esto suele ser una sorpresa para los legos en

matemáticas, a quienes se les enseñó en la escuela a comprobar si un número es primo por el mismo método que se utiliza para la factorización: ensayar todos los divisores posibles. El caso es que hay maneras ingeniosas para probar que un número es primo sin hacer eso. También nos permite demostrar que un número es compuesto sin encontrar ninguno de sus factores. Sencillamente mostrar que no supera un test de primalidad.

El bisabuelo de todos los tests modernos de primalidad es el teorema de Fermat, que no hay que confundir con el famoso último teorema de Fermat (véase capítulo 7). Este teorema se basa en la aritmética modular, a veces conocida como «aritmética de reloj» porque los números se enrollan como los de una esfera de reloj. Escogemos un número —para el caso análogo a un reloj de 12 horas es 12— y le llamamos el módulo. Ahora, en cualquier cálculo aritmético con números naturales se permite reemplazar cualquier múltiplo de 12 por cero. Por ejemplo, 5×5 = 25, pero 24 es dos veces 12, de modo que restando 24 obtenemos 5×5 = 1 módulo 12. La aritmética modular es muy bonita porque casi todas las reglas habituales de la aritmética siguen siendo válidas. La diferencia principal es que no siempre se puede dividir un número por otro, incluso si no es cero. La aritmética modular es también útil, porque proporciona una manera rigurosa de tratar cuestiones de divisibilidad: ¿qué números son divisibles por el módulo escogido, y cuál es el resto cuando no lo son? Gauss introdujo la aritmética modular en sus *Disquisitiones arithmeticae*, y hoy es ampliamente

utilizada en la ciencia de la computación, la física y la ingeniería, tanto como en las matemáticas.

El teorema de Fermat afirma que si escogemos un módulo primo p, y tomamos cualquier número a que no es un múltiplo de p, entonces la potencia (p - 1)-ésima de a es igual a 1 en aritmética módulo p. Supongamos, por ejemplo, que $p = 17$ y $a = 3$. Entonces el teorema predice que cuando dividimos 3^{16} por 17, el resto es 1. Como comprobación

$$3^{16} = 43.046.721 = 2.532.160 \times 17 + 1$$

Nadie en su sano juicio querría hacer las sumas de esta manera para primos de, digamos, cien dígitos. Afortunadamente, hay una manera más ingeniosa y rápida de realizar un cálculo de este tipo. Lo importante es que si la respuesta no es igual a 1, entonces el módulo con el que empezamos es compuesto. Por lo tanto, el teorema de Fermat constituye la base de un test eficaz que proporciona una condición necesaria para que un número sea primo.

Por desgracia, el test no es suficiente. Muchos números compuestos, conocidos como números de Carmichael, superan el test. El menor es 561, y en 2003 Red Alford, Andrew Granville y Carl Pomerance demostraron, para sorpresa general, que hay infinitos. Lo que era una sorpresa es que encontraron una demostración; el resultado real era menos sorprendente. De hecho,

ellos demostraron que hay al menos $x^{2/7}$ números de Carmichael menores o iguales que x si x es suficientemente grande.

Sin embargo, variantes más sofisticadas del teorema de Fermat pueden convertirse en auténticos testes de primalidad, tales como el publicado en 1976 por Gary Miller. Por desgracia, la prueba de validez del test de Miller depende de un gran problema no resuelto, la hipótesis de Riemann generalizada (véase capítulo 9). En 1980 Michael Rabin convirtió el test de Miller en un test probabilista, un test que en ocasiones podría dar la respuesta incorrecta. Las excepciones, si existen, son muy raras, pero no pueden descartarse por completo. El test determinista (es decir, que está garantizado que es correcto) más eficaz hasta ahora es el test de Adleman-Pomerance-Rumely, que debe su nombre a Leonard Adleman, Carl Pomerance y Robert Rumely. Utiliza ideas de la teoría de números que son más sofisticadas que el teorema de Fermat, pero con un espíritu similar.

Recuerdo muy bien una carta de un aficionado ilusionado que proponía una variante de una prueba de división. Ensayar todos los divisores posibles, pero empezando en la raíz cuadrada e ir *descendiendo*. Este método permite a veces llegar a la respuesta con más rapidez que haciendo las cosas en el orden habitual, pero cuando los números se hacen muy grandes tropieza con problemas similares a los del método habitual. Si lo ensayamos con mi ejemplo anterior, el número de 22 dígitos 1.080.813.321.843.836.712.253, entonces la raíz cuadrada es aproximadamente 32.875.725.419. Hay que ensayar 794.582.971 divisores primos antes de encontrar

uno que sea válido. Esto es *peor* que buscar en la dirección habitual.

En 1956 el famoso lógico Kurt Gödel, en un escrito a John von Neumann, se hacía eco de la petición de Gauss. Él preguntaba si la prueba de división podía mejorarse, y si es así, cuánto. Von Neumann no insistió en la cuestión, pero en el curso de los años otros respondieron a Gödel descubriendo métodos prácticos para encontrar primos de hasta cien dígitos, y a veces más. Estos métodos, de los que el más conocido es la llamada criba cuadrática, se han conocido desde aproximadamente 1980. Sin embargo, casi todos ellos son o bien probabilistas o bien son ineficaces en el sentido siguiente.

¿Cómo crece el tiempo de ejecución de un algoritmo computacional a medida que aumenta el tamaño del *input*? En el caso del test de primalidad, el tamaño del *input* no es el número en cuestión, sino cuántos dígitos tiene. La distinción esencial en tales cuestiones es la distinción entre dos clases de algoritmos llamados P y no-P. Si el tiempo de ejecución crece como alguna potencia dada del tamaño del *input*, entonces el algoritmo es de clase P; de lo contrario es no-P. En un sentido muy general, los algoritmos de clase P son útiles, mientras que los algoritmos no-P son poco prácticos, pero hay una tierra de nadie entre ambos donde entran en juego otras consideraciones. Aquí P representa «tiempo polinómico», una manera elegante de hablar de potencias, y volvemos al tema de los algoritmos eficaces en el capítulo 11.

Para la clase P estándar, la prueba por división funciona muy mal. Está bien en el aula, donde los números que aparecen tienen dos o tres dígitos, pero es del todo intratable para números de cien dígitos. La prueba por división está decididamente en la clase no-P. De hecho, el tiempo de ejecución se aproxima a 10 $n/2$ para un número de n dígitos, que crece con más rapidez que cualquier potencia de n. Un crecimiento de este tipo, llamado exponencial, es *realmente* malo, está en otro mundo computacional.

Hasta los años ochenta del siglo pasado todos los algoritmos conocidos para poner a prueba la primalidad, excluyendo los probabilistas o aquellos cuya validez no estaba demostrada, tenían un ritmo de crecimiento exponencial. Sin embargo, en 1983 se encontró un algoritmo que se halla tentadoramente en la tierra de nadie contigua al territorio P: al antes citado test de Adleman-Pomerance-Rumely. Una versión mejorada de Henri Cohen y Hendrik Lenstra tiene un tiempo de ejecución que va como n elevado a la potencia log log n, donde log denota el logaritmo. Técnicamente, log log n puede ser tan grande como queramos, de modo que este algoritmo no está en la clase P. Pero eso no le impide ser práctico: si n es un gúgolplex, un 1 seguido de 10^{100} ceros, entonces log log n es aproximadamente 230. Un viejo chiste dice: «Se ha demostrado que log log n tiende a infinito, pero nunca se ha observado que lo haga».

El primer test de primalidad de clase P fue descubierto en 2002 por Manindra Agrawal y sus estudiantes Neeraj Kayal y Nitin Saxena, quienes en aquella época todavía no se habían graduado. Doy

algunos detalles en la nota[10]. Ellos demostraron que su algoritmo tenía un tiempo de ejecución proporcional a lo sumo a n^{12}; esto fue mejorado rápidamente hasta $n^{7,5}$. Sin embargo, incluso si su algoritmo es de clase P, y por ello clasificado como «eficaz», sus ventajas no se ponen de manifiesto hasta que el número n se hace muy grande. Debería superar al test de Adleman-Pomerance-Rumely cuando el número de *dígitos* en n es aproximadamente 10^{1000}. No hay espacio en una memoria de ordenador o, de hecho, en todo el universo conocido, para incluir un número tan grande. Sin embargo, ahora que *sabemos* que existe un algoritmo de clase P para comprobar la primalidad, vale la pena buscar otros mejores. Lenstra y Pomerance redujeron el exponente de 7,5 a 6. Si son ciertas algunas otras conjeturas sobre los primos, entonces la potencia puede reducirse a 3, lo que empieza a parecer práctico.

Sin embargo, el aspecto más excitante del algoritmo de Agrawal-Kayal-Saxena no es el resultado sino el método. Es sencillo —para los matemáticos, al menos— y novedoso. La idea subyacente es una variante del teorema de Fermat, pero en lugar de trabajar con números, el equipo de Agrawal utilizó un polinomio. Este es una

---

[10] El algoritmo de Agrawal-Kayal-Saxena es como sigue:
Input: entero n
1. Si n es una potencia exacta de cualquier número más pequeño, resultado COMPUESTO y *stop*.
2. Encontrar el r más pequeño tal que la potencia más pequeña de r que sea igual a 1 módulo n es al menos $(\log n)^2$.
3. Si cualquier número menor o igual que r tiene un factor en común con n, resultado COMPUESTO y *stop*.
4. Si n es menor o igual que r, resultado PRIMO y *stop*.
5. Para todos los números naturales a desde 1 hasta un límite especificado, comprobar si el polinomio $(x + a)^n$ es igual a $x^n + a$, módulo n y módulo $x^r - 1$. Si la igualdad se cumple en algún caso, resultado COMPUESTO y *stop*.
6. Resultado PRIMO.

combinación de potencias de una variable x, tal como $5x^3 + 4x - 1$. Se pueden sumar, restar y multiplicar polinomios, y las reglas algebraicas usuales siguen siendo válidas. El capítulo 3 explica los polinomios con más detalle.

Esta es una idea realmente hermosa: extender el dominio del discurso y trasladar el problema a un nuevo ámbito de pensamiento. Es una de esas ideas que son tan simples que hay que ser un genio para descubrirlas. Deriva de un artículo de 1999 de Agrawal y su director de tesis Somenath Biswas, que daba un test de primalidad probabilista basado en un análogo del teorema de Fermat en el mundo de los polinomios. Agrawal estaba convencido de que el elemento probabilista podía eliminarse. En 2001 sus estudiantes dieron con una observación crucial y bastante técnica. Seguir esa idea llevó al equipo a aguas profundas en teoría de números, pero finalmente todo se redujo a un único obstáculo, la existencia de un primo p tal que $p - 1$ tiene un divisor primo suficientemente grande. Algunas consultas y búsquedas en internet llevaron a un teorema demostrado por Etienne Fouvry en 1985 utilizando métodos profundos y técnicos. Esto era exactamente lo que necesitaban para demostrar que su algoritmo funcionaba, y la pieza final del rompecabezas encajó limpiamente en su lugar.

En los días en que la teoría de números estaba encerrada dentro de su propia torre de marfil, nada de esto habría importado al resto del mundo. Pero durante los últimos veinte años, los números primos se han hecho importantes en criptografía, la ciencia de los códigos secretos. Los códigos no solo son importantes para uso militar,

también las empresas comerciales tienen secretos. En esta era de internet, todos nosotros los tenemos: no queremos que los criminales tengan acceso a nuestras cuentas bancarias, números de tarjetas de crédito o, con el aumento de robos de identidad, el nombre de nuestro gato. Pero internet es una forma tan conveniente de pagar facturas, asegurar automóviles y contratar vacaciones, que tenemos que aceptar cierto riesgo de que nuestra información privada y sensible pueda caer en manos equivocadas.

Los fabricantes de ordenadores y los proveedores de internet tratan de reducir el riesgo haciendo disponibles varios sistemas de encriptación. La implicación de los ordenadores ha transformado la criptografía y el criptoanálisis, el oscuro arte de romper códigos. Se han ideado muchos códigos novedosos, y uno de los más famosos, inventado por Ron Rivest, Adi Shamir y Leonard Adleman en 1978, utiliza números primos. Números grandes, de unos cien dígitos. El sistema Rivest-Shamir-Adleman se utiliza en los principales protocolos para comunicación segura en internet, y es ampliamente utilizado por gobiernos, corporaciones y universidades. Eso no significa que todo resultado nuevo sobre primos sea importante para la seguridad de su cuenta bancaria en internet, pero añade cierta excitación a cualquier descubrimiento que relacione los primos con la computación. El test de Agrawal-Kayal-Saxena es un ejemplo al respecto. Matemáticamente es elegante e importante, pero no tiene trascendencia práctica directa.

Sin embargo, arroja una nueva y ligeramente perturbadora luz sobre la cuestión general de la criptografía de Rivest-Shamir-

Adleman. Todavía no hay ningún algoritmo de clase P para resolver el segundo problema de Gauss, la factorización. La mayoría de los expertos piensan que no existe nada de ese tipo, pero no están tan seguros como solían estarlo. Puesto que nuevos descubrimientos como el test de Agrawal-Kayal-Saxena pueden aguardar escondidos en cualquier rincón, basados en ideas tan simples como las versiones polinómicas del teorema de Fermat, los criptosistemas basados en la factorización en primos podrían no ser tan seguros como orgullosamente imaginamos. No revele todavía el nombre de su gato en internet.

Incluso las matemáticas básicas de los números primos llevan rápidamente a conceptos más avanzados. El misterio se hace aún más profundo cuando planteamos preguntas más sutiles. Euclides demostró que los primos no tienen fin, de modo que no podemos hacer una lista de ellos y detenernos. Ni tampoco podemos dar una fórmula algebraica simple y útil para primos sucesivos, de forma similar a como x^2 especifica cuadrados. (Existen fórmulas simples, pero «hacen trampa» al incorporar en la fórmula a los primos disfrazados, y no nos dicen nada nuevo.[11]) Para captar la naturaleza de estos números elusivos y erráticos, podemos realizar experimentos, buscar indicios de estructura y tratar de demostrar

[11] Un ejemplo de lo que tengo en mente es la fórmula $[A^{3^n}]$, donde los corchetes denotan el máximo entero menor o igual que su contenido. En 1947 W. H. Mills demostró que existe una constante real A tal que esta fórmula da un primo para cualquier n. Suponiendo la hipótesis de Riemann, el mínimo valor de A que funciona es aproximadamente 1,306. Sin embargo, la constante está definida utilizando una secuencia apropiada de primos, y la fórmula es tan solo una forma simbólica de reproducir esta secuencia. Para más de estas fórmulas, incluidas algunas que representan a todos los primos, véase http://mathworld.wolfram.com/PrimeFormulas.html
http://en.wikipedia.org/wiki/Formula_for_primes.

que estas pautas aparentes persisten por muy grandes que se hagan los primos. Por ejemplo, podemos preguntar cómo se distribuyen los primos entre los números naturales. Las tablas de números primos sugieren con fuerza que ellos tienden a hacerse más raros a medida que se hacen más grandes. La Tabla 1 muestra cuántos primos hay en varios intervalos de 1000 números consecutivos.

TABLA 1. Número de primos en intervalos sucesivos de 1000 números.

Rango	*Número de primos*
1-1.000	168
1.001-2.000	135
2.001-3.000	127
3.001-4.000	119
4.001-5.000	118
5.001-6.000	114
6.001-7.000	117
7.001-8.000	106
8.001-9.000	110
9.001-10.000	111

Los números en la segunda columna decrecen en general a medida que descendemos por la tabla, aunque a veces hay breves períodos en que varían en sentido contrario: por ejemplo, 114 va seguido por 117. Esto es un síntoma de la irregularidad de los primos, pero a pesar de eso hay una clara tendencia general de los primos a hacerse más raros a medida que su tamaño aumenta. La razón no es difícil de ver: cuanto mayor se hace un número, más factores potenciales existen. Los primos tienen que evitar todos estos

factores. Es como pescar no primos con una red: cuanto más fina se hace la red, menos primos se escapan de ella.

La «red» incluso tiene un nombre: la criba de Eratóstenes. Eratóstenes de Cirene fue un matemático griego que vivió alrededor de 250 a. C. Fue también un atleta con intereses en poesía, geografía, astronomía y música. Hizo la primera estimación razonable del tamaño de la Tierra observando la posición del Sol al mediodía en dos lugares diferentes, Alejandría y Siena (la actual Asuán). Al mediodía el Sol estaba exactamente en la vertical de Siena, pero a unos 7 grados de la vertical en Alejandría. Puesto que este ángulo es una cincuentava parte de un círculo, la circunferencia de la Tierra debe ser cincuenta veces la distancia entre Alejandría y Siena. Eratóstenes no podía medir esta distancia directamente, de modo que preguntó a los mercaderes cuánto tiempo tardaban en hacer el viaje en camello y estimó cuánto recorría típicamente un camello en un día. Dio una cifra explícita en una unidad conocida como *stadium*, pero no sabemos cuánto valía dicha unidad.

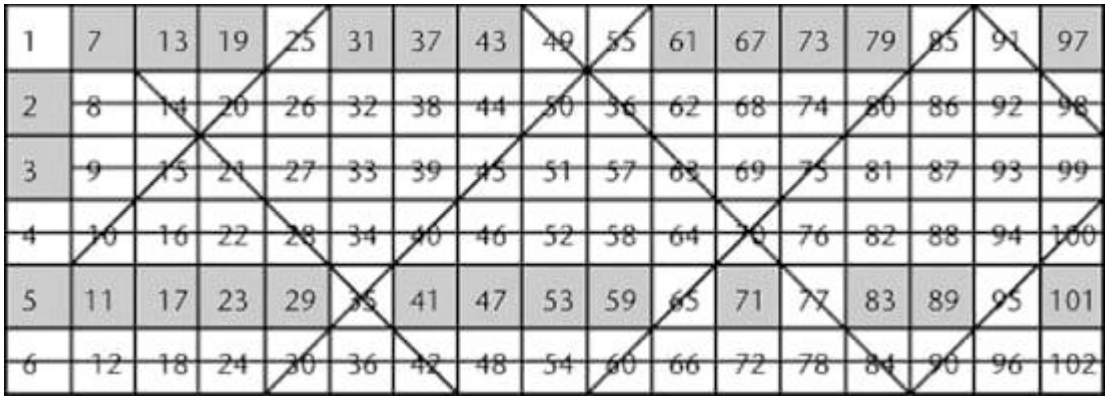


Figura 2. La criba de Eratóstenes.

Los historiadores piensan por lo general que la estimación de Eratóstenes era razonablemente aproximada.

Su criba es un algoritmo para encontrar todos los primos eliminando de modo sucesivo todos los múltiplos de números que ya se sabe que son primos. La Figura 2 ilustra el método sobre los números hasta 102, dispuestos para hacer el proceso de eliminación fácil de seguir. Para ver cómo procede, le sugiero que construya el diagrama usted mismo. Empiece solo con la malla, omitiendo las líneas que cruzan los números. Luego puede añadir dichas líneas una a una. Omita 1 porque es una unidad. El siguiente número es 2, de modo que es primo. Cruce todos los múltiplos de 2: estos se encuentran en las líneas horizontales que parten de 4, 6 y 8. El siguiente número no cruzado es 3, de modo que es primo. Cruce todos los múltiplos de 3: se encuentran en las líneas horizontales que parten de 6, ya cruzados, y 9. El siguiente número no cruzado es 5, de modo que es primo. Cruce todos los múltiplos de 5: se encuentran en las líneas diagonales inclinadas hacia arriba y la derecha, partiendo de 10. El siguiente número no cruzado es 7, de modo que es primo. Cruce todos los múltiplos de 7: se encuentran en las líneas diagonales inclinadas hacia abajo y la derecha, partiendo de 14. El siguiente número no cruzado es 11, de modo que es primo. El primer múltiplo de 11 que no ha sido ya cruzado porque tiene un divisor más pequeño es 121, que está fuera de la imagen, de modo que nos detenemos. Los números que quedan sombreados, son los primos.

La criba de Eratóstenes no es solo una curiosidad histórica; sigue siendo uno de los métodos más efectivos que se conoce para hacer listas extensas de primos. Y métodos relacionados han llevado a progresos importantes sobre lo que probablemente es el más famoso problema no resuelto sobre los primos: la conjetura de Goldbach. El matemático aficionado alemán Christian Goldbach mantenía correspondencia con muchas de las figuras famosas de su época. En 1742 enunció varias curiosas conjeturas sobre primos en una carta a Leonhard Euler. Los historiadores advirtieron más tarde que René Descartes había dicho prácticamente lo mismo algunos años antes. El primero de los enunciados de Goldbach era: «Todo número entero que puede escribirse como suma de dos primos, puede escribirse también como suma de tantos primos como se quiera, hasta que todos los términos sean unidades». El segundo, añadido en el margen de su carta, era: «Todo entero mayor que 2 puede escribirse como suma de tres primos». Con la definición actual de «primo» hay excepciones obvias a estos enunciados. Por ejemplo, 4 no es suma de tres primos, porque el primo más pequeño es 2, de modo que la suma de tres primos debe ser al menos 6. Pero en los días de Goldbach el número 1 se consideraba primo. Es sencillo reenunciar sus conjeturas utilizando el convenio moderno.

En su respuesta, Euler recordaba una conversación anterior con Goldbach, cuando este había señalado que su primera conjetura se seguía de una más sencilla, su tercera conjetura: «Todo número par es suma de dos primos». Con el convenio imperante de que 1 es primo, este enunciado implica también la segunda conjetura,

porque cualquier número puede escribirse como $n + 1$ o como $n + 2$ donde n es par. Si n es la suma de dos primos, el número original es suma de tres primos. La opinión de Euler acerca de la tercera conjetura era inequívoca: «La considero como un teorema absolutamente cierto, aunque no puedo demostrarlo». Esto resume a la perfección su estatus actual.

El convenio moderno, en el que 1 no es primo, separa las conjeturas de Goldbach en dos diferentes. La conjetura de Goldbach par afirma:

Todo entero par mayor que dos es suma de dos primos.

La conjetura de Goldbach impar es:

Todo número impar mayor que 5 es suma de tres primos.

La conjetura par implica la impar, pero no a la inversa[12]. Es útil considerar ambas conjeturas por separado porque aún no sabemos si alguna de ellas es cierta. La conjetura impar parece ser ligeramente más fácil que la par, en el sentido de que se han hecho más progresos.

Algunos cálculos rápidos verifican la conjetura de Goldbach par para números pequeños.

$$4 = 2 + 2$$

$$6 = 3 + 3$$

$$8 = 5 + 3$$

$$10 = 7 + 3 = 5 + 5$$

[12] Si n es impar entonces $n - 3$ es par, y si n es mayor que 5 entonces $n - 3$ es mayor que 2. Por la primera conjetura, $n - 3 = p + q$, de modo que $n = p + q + 3$.

$$12 = 7 + 5$$

$$14 = 11 + 3 = 7 + 7$$

$$16 = 13 + 3 = 11 + 5$$

$$18 = 13 + 5 = 11 + 7$$

$$20 = 17 + 3 = 13 + 7$$

Es fácil continuar a mano hasta, digamos, 1000 o más si uno es persistente. Por ejemplo 1000 = 3 + 997, y 1.000.000 = 17 + 999.993. En 1938 Nils Pipping verificó la conjetura de Goldbach par para todos los números pares hasta 100.000.

También se hizo evidente que a medida que el número en cuestión se hace más grande, tiende a haber cada vez más maneras de escribirlo como suma de primos. Esto tiene sentido. Si se toma un número par grande, y se siguen restando primos uno tras otro, ¿qué probabilidad hay de que *todos* los resultados sean compuestos? Basta con que aparezca un primo entre la lista de diferencias resultante para que la conjetura quede verificada para dicho número. Utilizando las propiedades estadísticas de los primos, podemos evaluar la probabilidad de un resultado semejante. Los analistas Godfrey Harold Hardy y John Littlewood realizaron un cálculo semejante en 1923, y derivaron una fórmula plausible aunque no rigurosa para el número de maneras diferentes de expresar un número par dado n como suma de dos primos: hay aproximadamente $n/[2(\log n)^2]$ maneras. Este número aumenta a medida que n se hace más grande, y también está de acuerdo con la evidencia numérica. Pero incluso si este cálculo pudiera hacerse

riguroso, podría ser solamente una rara y ocasional excepción, así que no es de gran ayuda.

El principal obstáculo para una demostración de la conjetura de Goldbach es que combina dos propiedades muy diferentes. Los primos están definidos en términos de multiplicación, pero las conjeturas se refieren a la suma. Por eso es extraordinariamente difícil relacionar la conclusión deseada con cualesquiera propiedades razonables de los primos. Parece que no hay ningún lugar para apoyar una palanca. Debe haber sido música para los oídos de la editorial Faber & Faber en 2000, cuando ofreció un premio de un millón de dólares por una demostración de la conjetura para promocionar la novela *El tío Petros y la conjetura de Goldbach* de Apostolos Doxiadis. El plazo era rígido: había que presentar una solución antes de abril de 2002. Nadie reclamó el premio, lo que apenas sorprende dado que el problema ha permanecido sin resolver durante más de 250 años.

La conjetura de Goldbach suele reformularse como una cuestión acerca de sumar conjuntos de números enteros. La conjetura de Goldbach par es el ejemplo más simple de esta particular forma de pensar, porque sumamos precisamente *dos* conjuntos de enteros. Para hacerlo, tomamos cualquier número del primer conjunto, le sumamos cualquier número del segundo conjunto, y luego tomamos el conjunto de todas estas sumas. Por ejemplo, la suma de {1, 2, 3} y {4, 5} contiene 1 + 4, 2 + 4, 3 + 4, 1 + 5, 2 + 5, 3 + 5, que es {5, 6, 7, 8}. Algunos números aparecen más de una vez, por ejemplo 6 = 2 + 4 = 1 + 5. Llamaré «solapamiento» a este tipo de repetición.

La conjetura de Goldbach par puede reenunciarse ahora: si sumamos el conjunto de primos consigo mismo, el resultado contiene todo número par mayor que 2. Esta reformulación puede sonar un poco trillada —y lo es— pero introduce el problema en un área donde existen algunos potentes teoremas generales. El número 2 es una pequeña molestia, pero podemos prescindir de él con facilidad. Es el único primo par, y si lo sumamos a cualquier otro primo el resultado es impar. De modo que por lo que respecta a la conjetura de Goldbach par podemos olvidarnos del 2. Sin embargo, necesitamos 2 + 2 para representar 4, por lo que también debemos restringir la atención a números pares que sean al menos 6.

A modo de experimento sencillo consideremos los números pares hasta 30, incluido. Hay nueve primos impares en este rango: {3, 5, 7, 11, 13, 17, 19, 23, 29}. Sumarlos da la Figura 3: he marcado en negrita las sumas que son menores o iguales que 30 (un rango de números pares que incluye todos los primos hasta 29). Aparecen dos pautas simples. La tabla entera es simétrica respecto a su diagonal principal puesto que $a + b = b + a$. Los números en negrita ocupan aproximadamente la mitad superior izquierda de la tabla, sobre la línea (diagonal) gruesa. Si acaso, tienden a rebasarla en el centro. Esto sucede porque, en conjunto, los primos más grandes son más raros que los pequeños. La región extra que la rebasa compensa con creces los dos 32 arriba a la derecha y abajo a la izquierda.

	3	5	7	11	13	17	19	23	29
3	6	8	10	14	16	20	22	26	32
5	8	10	12	16	18	22	24	28	34
7	10	12	14	18	20	24	26	30	36
11	14	16	18	22	24	28	30	34	40
13	16	18	20	24	26	30	32	36	42
17	20	22	24	28	30	34	36	40	46
19	22	24	26	30	32	36	38	42	48
23	26	28	30	34	36	40	42	46	52
29	32	34	36	40	42	46	48	52	58

Figura 3. Sumas de pares de primos hasta 30. En negrita: sumas que son 30 o menor. Línea gruesa: diagonal. Región sombreada: se eliminan pares simétricamente relacionados. La región sombreada ocupa algo más de una cuarta parte del cuadrado.

Ahora hacemos algunas crudas estimaciones. Podría ser más preciso, pero estas son suficientemente buenas. El número de casillas en la tabla es 9×9 = 81. Alrededor de la mitad de los números en dichas casillas están en el triángulo superior izquierdo. Debido a la simetría aparecen en pares excepto a lo largo de la diagonal, de modo que el número de casillas no emparentadas es más o menos 81/4, aproximadamente 20. El número de enteros pares en el rango de 6 a 30 es 13. Así que las 20 (y más) sumas en negrita tienen que dar solo 13 números pares. Hay más sumas potenciales de dos primos en el rango correcto que números pares.

Es como arrojar 20 bolas a 13 cocos en la feria. Uno tiene una probabilidad razonable de dar en un montón de ellos. Incluso así, uno podría dejar de dar en algunos cocos. Algunos números pares podrían seguir faltando.

En este caso no lo hacen, pero este tipo de argumento de recuento no puede eliminar esa posibilidad. Sin embargo, nos dice que debe haber bastante más de un solapamiento, donde el mismo número en negrita aparece varias veces en el cuadrante relevante de la tabla. ¿Por qué? Porque 20 sumas tienen que encajar en un conjunto con solo 13 miembros. Así que en promedio cada número en negrita aparece aproximadamente 1,5 veces. (El número real de sumas es 27, de modo que una mejor estimación muestra que cada número en negrita aparece dos veces). Si faltan algunos números pares, el solapamiento debe ser aún mayor.

Podemos jugar a lo mismo con un límite superior más grande —digamos un millón—. Una fórmula llamada el teorema de los números primos (véase capítulo 9), proporciona una simple estimación del número de primos hasta cualquier tamaño x dado. La fórmula es $x/\log x$. En este caso, la estimación es de aproximadamente 72.380. (La cifra exacta es 78.497). La región sombreada correspondiente ocupa aproximadamente un cuarto de la tabla, de modo que proporciona alrededor de $n^2/4 = 250.000$ millones de números en negrita: sumas de dos primos en este rango. Esto es inmensamente mayor que el número de números pares en el rango, que es de medio millón. Ahora la cantidad de solapamiento tiene que ser gigantesca, con cada suma apareciendo

en promedio quinientas mil veces. De modo que la probabilidad de que se escape cualquier número par concreto se reduce enormemente.

Con más esfuerzo podemos convertir esta aproximación en una estimación de la probabilidad de que algún número par en un rango dado no sea la suma de dos primos, suponiendo que los primos estén distribuidos al azar y con frecuencias dadas por el teorema de los números primos, es decir, en torno a $x/\log x$ primos menores que cualquier x dado. Esto es lo que hicieron Hardy y Littlewood. Ellos sabían que su aproximación no era rigurosa, porque los primos se definen por un proceso específico y no son realmente aleatorios. No obstante, es razonable esperar que los resultados reales sean consistentes con este modelo probabilista, porque la propiedad definitoria de los primos parece tener muy poca relación con lo que sucede cuando sumamos dos de ellos.

Varios métodos estándar en esta área adoptan un punto de vista similar, aunque poniendo un cuidado extra para hacer el argumento riguroso. Ejemplos de ellos son los métodos de criba, que se basan en la criba de Eratóstenes. Los teoremas generales sobre la densidad de números en sumas de dos conjuntos —la proporción de números que aparecen a medida que los conjuntos se hacen muy grandes— suministran otras herramientas útiles.

Cuando finalmente una conjetura matemática resulta ser correcta, su historia suele seguir una pauta estándar. Durante un período de tiempo varias personas demuestran que la conjetura es correcta siempre que se apliquen restricciones especiales. Cada uno de estos

resultados mejora los anteriores al relajar algunas restricciones, pero con el tiempo este proceso pierde fuelle. Por último, una idea nueva y mucho más ingeniosa completa la demostración.

Por ejemplo, una conjetura en teoría de números puede afirmar que todo entero positivo puede representarse de cierta manera utilizando, digamos, seis números especiales (primos, cuadrados, cubos, lo que sea). Aquí las características clave son *todo* entero positivo y *seis* números especiales. Avances iniciales llevan a resultados mucho más débiles, pero sucesivas etapas en el proceso los mejoran lentamente.

El primer paso suele ser una demostración que sigue estas líneas: todo entero positivo que no es divisible por 3 o por 11, excepto un número finito de ellos, puede representarse en términos de un número gigantesco de números especiales, digamos 10^{666}. Normalmente el teorema no especifica cuántas excepciones hay, de modo que el resultado no puede aplicarse directamente a un entero concreto. El siguiente paso es hacer la cota efectiva: es decir, demostrar que todo entero mayor que 10^{1042} puede representarse así. Luego se elimina la restricción a la divisibilidad por 3, seguida por un avance similar para 11. Después de eso, sucesivos autores reducen uno de los números 10^{666} o 10^{1042}, a veces ambos. Una típica mejora podría ser que todo entero mayor que $5{,}8\times10^{17}$ puede representarse utilizando como máximo 4298 números especiales, por ejemplo.

Mientras, otros investigadores están trabajando de abajo arriba partiendo de números pequeños, a menudo con ayuda de un

ordenador, demostrando que, digamos, todo número menor o igual que 10^{12} puede representarse utilizando a lo sumo seis números especiales. En menos de un año, 10^{12} ha sido mejorado en cinco etapas, por diferentes investigadores o grupos, hasta $11{,}0337\times10^{29}$. Estas mejoras no son rutinarias ni fáciles, pero la forma en que se consiguen incluye complicados métodos especiales que no proporcionan ningún indicio de una aproximación más general, y cada contribución sucesiva es más complicada y más larga. Después de algunos años de este tipo de pequeñas mejoras, aplicando las mismas ideas generales pero con ordenadores más potentes y nuevos pequeños retoques, este número se ha elevado hasta 10^{43}. Pero ahora los métodos están en un punto muerto, y todo el mundo está de acuerdo en que por muchos retoques que se hagan nunca se llegará a la conjetura completa.

En este momento la conjetura desaparece de la vista porque nadie está trabajando ya en ella. A veces, el progreso se para en seco. Otras veces pasan veinte años sin nada nuevo... y luego, aparentemente de la nada, Cheesburger y Fries anuncian que reformulando la conjetura en términos de cuasicúmulos metaergódicos complejos y aplicando teoría colaboracionista bizantina, han obtenido una demostración completa. Tras varios años discutiendo sobre detalles lógicos, y llenar unas pocas lagunas, la comunidad matemática acepta que la demostración es correcta, e inmediatamente pregunta si hay un modo mejor de obtener el mismo resultado o llevarlo aún más lejos.

En capítulos posteriores veremos cómo esta misma pauta se repite muchas veces. Puesto que tales exposiciones se hacen tediosas, por muy orgullosos que estén Buggins y Krumm de su más reciente mejora del exponente en la conjetura de Jekyll-Hyde de 1,773 a 1,771 + ε para cualquier ε positivo, yo describiré unas pocas contribuciones representativas y dejaré fuera las demás. Esto no es negar la importancia del trabajo de Buggins y Krumm. Incluso es posible que haya preparado el camino para el logro trascendental de Cheesburger-Fries. Pero es probable que solo los expertos, que siguen el desarrollo de la historia, esperen la próxima mejora minúscula conteniendo la respiración.

En el futuro proporcionaré menos detalles, pero veamos cómo es en el caso de Goldbach.

Se han establecido teoremas que recorren algún camino hacia la demostración de la conjetura de Goldbach. El primer gran avance llegó en 1923, cuando Hardy y Littlewood utilizaron sus técnicas analíticas para demostrar la conjetura de Goldbach impar para todos los números impares suficientemente grandes. Sin embargo, su demostración se basaba en otra gran conjetura, la hipótesis de Riemann generalizada, que se discute en el capítulo 9. Este problema sigue abierto, de modo que su aproximación tenía una laguna importante. En 1930 Lev Schnirelmann salvó la laguna utilizando una versión imaginativa del razonamiento de aquellos, basada en métodos de criba. Él demostró que una proporción no nula de todos los números puede representarse como una suma de dos primos. Combinando este resultado con algunas generalidades

sobre sumas de secuencias, demostró que hay un número C tal que todo entero mayor que 1 es una suma de como máximo C números primos. Este número llegó a conocerse como constante de Schnirelmann. Ivan Matveyevich Vinogradov obtuvo resultados similares en 1937, pero su método tampoco especificaba cómo de grande es «significativamente grande». En 1939 K. Borozdin demostró que no es mayor que $3^{14.348.907}$. Para 2002 Liu Ming-Chit y Wang Tian-Ze habían reducido esta «cota superior» a e^{3100}, que es aproximadamente 2×10^{1346}. Esto es mucho más pequeño, pero sigue siendo demasiado grande para que los números intermedios sean comprobados por ordenador.

En 1969 N. I. Klimov obtuvo la primera estimación específica para la constante de Schnirelmann: es a lo sumo seis mil millones. Otros matemáticos redujeron ese número considerablemente, y para 1982 Hans Riesel y Robert Vaughan lo habían reducido a 19. Aunque 19 es mucho mejor que seis mil millones, la evidencia apuntaba a que la constante de Schnirelmann era un simple 3. En 1995 Leszek Kaniecki redujo la cota superior a 6, con cinco primos para cualquier número impar, pero tuvo que suponer que la hipótesis de Riemann es verdadera. Sus resultados, combinados con la verificación numérica que hizo J. Richstein de la hipótesis de Riemann hasta 4×10^{14}, demostrarían que la constante de Schnirelmann es a lo sumo 4, suponiendo de nuevo la hipótesis de Riemann. En 1997 Jean-Marc Deshouillers, Gove Effinger, Herman te Riele y Dimitrii Zinoviev demostraron que la hipótesis de Riemann generalizada (véase capítulo 9) implica la conjetura de Riemann

impar. Es decir, todo número impar excepto 1, 3 y 5 es la suma de tres primos.

Puesto que la hipótesis de Riemann no está demostrada actualmente, vale la pena tratar de remover esta hipótesis. En 1995 el matemático francés Olivier Ramaré redujo la estimación superior para representar números impares hasta 7, sin utilizar la hipótesis de Riemann. De hecho, él demostró algo más fuerte: todo número par es una suma de a lo sumo seis números primos. (Para tratar con números impares, restar 3: el resultado es par, de modo que es una suma de seis o menos primos. El número original es esta suma más el primo 3, lo que requiere siete o menos primos). El avance principal fue mejorar las estimaciones existentes para la proporción de números, en un rango especificado, que son suma de dos primos. El resultado clave de Ramaré es que para cualquier número n mayor que e^{67} (aproximadamente $1{,}25\times10^{29}$), al menos una quinta parte de los números entre n y $2n$ son suma de dos primos. Utilizando métodos de cribas, junto con un teorema de Hans-Heinrich Ostmann sobre sumas de secuencias, refinado por Deshouillers, esto lleva a una demostración de que todo número par mayor que 10^{30} es una suma de a lo sumo seis primos.

El obstáculo que queda es tratar el hueco entre 4×10^{14}, donde Jörg Richstein había comprobado el teorema por ordenador, y 10^{30}. Como es habitual, los números son demasiado grandes para una búsqueda directa por ordenador, de modo que Ramaré demostró una serie de teoremas especializados sobre el número de primos en intervalos pequeños. Estos teoremas dependen de la verdad de la

hipótesis de Riemann hasta límites específicos, que puede ser verificada por ordenador. Por ello, la prueba consiste principalmente en deducciones conceptuales de papel y lápiz, con ayuda de ordenador en este aspecto concreto. Ramaré terminaba su artículo señalando que en principio una aproximación similar podría reducir el número de primos de siete a cinco. Sin embargo, había enormes obstáculos prácticos, y escribió que una demostración semejante «no puede alcanzarse con los ordenadores actuales».

En 2012 Terence Tao superó estas dificultades con algunas ideas nuevas y muy diferentes. Él colocó un artículo en internet, que en el momento en que yo escribo está bajo revisión para su publicación. Su teorema principal es: todo número impar es una suma de a lo sumo cinco primos. Esto reduce la constante de Schnirelmann a 6. Tao es reputado por su capacidad para resolver problemas difíciles en muchas áreas de las matemáticas. Su demostración introduce varias técnicas potentes en el problema y requiere ayuda del ordenador. Si el número 5 en el teorema de Tao pudiera reducirse a 3, la conjetura de Goldbach impar estaría demostrada, y la cota sobre la constante de Schnirelmann se reduciría a 4. Tao sospecha que sería posible hacerlo, aunque se necesitarán más nuevas ideas.

La conjetura de Goldbach par parece aún más difícil. En 1998 Deshouillers, Saouter y Te Riele la verificaron para todos los números pares hasta 10^{14}. Para 2007, Tomás Oliveira e Silva lo había mejorado hasta 10^{18}, y sus computaciones continúan. Sabemos que todo entero par es la suma de a lo sumo seis primos —demostrado por Ramaré en 1995—. En 1973 Chen Jing-Run

demostró que todo entero par suficientemente grande es la suma de un primo y un semiprimo (o bien un primo o un producto de dos primos). Esto se acerca, pero todavía no hay premio. Tao ha afirmado que la conjetura de Goldbach par está más allá del alcance de sus métodos. Sumar tres primos crea mucho más solapamiento en los números resultantes —en el sentido discutido en relación con la Figura 3— que los dos primos necesarios para la conjetura de Goldbach par, y los métodos de Tao y Ramaré explotan repetidamente esta propiedad.

Quizá, entonces, en pocos años tengamos una demostración completa de la conjetura de Goldbach impar, que implica en particular que todo número impar es la suma de a lo sumo cuatro primos. Pero la conjetura de Goldbach par seguirá siendo probablemente tan desconcertante como lo fue para Euler y Goldbach.

En los 2300 años transcurridos desde que Euclides demostró varios teoremas básicos sobre los primos hemos aprendido mucho sobre estos números evasivos aunque de vital importancia. Pero lo que ahora sabemos da una cruda perspectiva de la larga lista de lo que no sabemos.

Sabemos, por ejemplo, que hay infinitos primos de la forma $4k + 1$ y $4k + 3$; con más generalidad, que cualquier secuencia aritmética[13] $ak + b$ para a y b dados contiene infinitos primos siempre que a y b no tengan ningún factor común. Por ejemplo, supongamos que $a = 18$. Entonces $b = 1, 5, 7, 11, 13$ o 17. Por consiguiente, existen

[13] Prefiero este término al antiguo, pero quizá más familiar, «progresión aritmética». Nadie habla ya de progresiones, excepto en los casos de la aritmética y la geométrica.

infinitos primos de cada una de las formas $18k + 1$, $18k + 5$, $18k + 7$, $18k + 11$, $18k + 13$ o $18k + 17$. Esto no es cierto para, digamos, $18k + 6$, porque esto es un múltiplo de 6. Ninguna secuencia aritmética puede contener primos *solamente*, pero un reciente avance importante, el teorema de Green-Tao, muestra que el conjunto de primos contiene secuencias aritméticas arbitrariamente largas. La demostración, obtenida en 2004 por Ben Green y Terence Tao, es profunda y difícil. Nos da esperanzas: difíciles cuestiones abiertas, por impenetrables que puedan parecer, pueden a veces ser respondidas.

Poniéndonos el sombrero de algebristas nos preguntamos inmediatamente por fórmulas más complicadas que incluyen k. No hay primos de la forma k^2, y ninguno salvo 3 de la forma $k^2 - 1$, porque estas expresiones factorizan. Sin embargo, la expresión $k^2 + 1$ no tiene factores obvios, y aquí podemos encontrar muchos primos:

$$2 = 1^2 + 15 = 2^2 + 117 = 4^2 + 137 = 6^2 + 1$$

y así sucesivamente. Un ejemplo mayor sin significado especial es:

$$18.672.907.718.657 = (4.321.216)^2 + 1$$

Se conjetura que existen infinitos de tales primos, pero todavía no se ha demostrado ninguna afirmación semejante para ningún polinomio específico en el que aparezca k elevado a una potencia

superior a la primera. Una conjetura muy plausible es la propuesta por V. Bouniakowsky en 1857: cualquier polinomio en k que no tenga divisores obvios representa infinitos primos. Las excepciones a esto incluyen no solo polinomios reducibles, sino polinomios como $k^2 + k + 2$ que siempre es divisible por 2, pese a no tener factores algebraicos.

Algunos polinomios parecen tener propiedades especiales. El caso clásico es $k^2 + k + 41$, que es primo para $k = 0, 1, 2..., 40$, y de hecho también para $k = -1, -2..., -40$. Largas ristras de primos para valores consecutivos de k son raras, y se tienen algunos conocimientos sobre ellas. Pero el área general es muy misteriosa.

Casi tan famosa como la conjetura de Goldbach, y aparentemente tan difícil como esta, es la conjetura de los primos gemelos: hay infinitos pares de primos que difieren en 2. Ejemplos de ellos son:

3, 5; 5, 7; 11,13; 17, 19

Los mayores primos gemelos conocidos (hasta enero de 2012) son:

$$3.756.801.695.685 \times 2^{666.669} \pm 1$$

que tienen 200.700 cifras decimales. Fueron encontrados en 2011 por el proyecto de computación distribuida PrimeGrid. En 1915, Viggo Brun utilizó una variante de la criba de Eratóstenes para demostrar que la suma de los recíprocos de todos los primos gemelos converge, a diferencia de la suma de los recíprocos de todos

los primos. De modo que, en este sentido, los primos gemelos son relativamente raros. Él también demostró, utilizando métodos similares, que existen infinitos enteros n tales que n y $n + 2$ tienen a lo sumo nueve factores primos. Hardy y Littlewood utilizaron sus métodos heurísticos para argumentar que el número de pares de primos gemelos menores que x debería tender asintóticamente a:

$$2a\frac{n}{(\log n)^2}$$

donde a es una constante cuyo valor es aproximadamente 0,660161. La idea subyacente es que para este fin puede suponerse que los primos se dan al azar, a un ritmo que hace que el número de primos hasta x es aproximadamente igual a $x/\log x$. Hay muchas conjeturas y fórmulas heurísticas similares pero, una vez más, no hay demostraciones rigurosas.

De hecho, hay cientos de cuestiones abiertas sobre los primos. Algunas son tan solo curiosas, otras son profundas e importantes. Encontraremos algunas de las últimas en el capítulo 9. Pese a todos los avances que los matemáticos han hecho durante los últimos 2500 años, los humildes primos no han perdido nada de su atractivo ni nada de su misterio.

Capítulo 3

El rompecabezas de pi.

La cuadratura del círculo

Los primos son una idea vieja, pero los círculos son todavía más dos mil años en resolverse. Es uno de varios problemas geométricos relacionados que nos han llegado de la Antigüedad. El personaje central en la historia es el número π (el «pi» griego) que encontramos en la escuela en relación con los círculos y las esferas. Su valor numérico es 3,14159 y un poco más; a veces se utiliza la aproximación 22/7. Las cifras decimales de π no se acaban nunca, y nunca repiten la misma secuencia una y otra vez. El récord actual de cifras calculados de π lo ostentan Alexander Yee y Chigeru Kondo que calcularon un billón de cifras en octubre de 2011[14]. Este tipo de computaciones es importante como manera de poner a prueba ordenadores rápidos, o para inspirar y probar nuevos métodos de cálculo de π, pero muy poco depende de los resultados numéricos. La razón para interesarnos en π no es calcular la longitud de una circunferencia. El mismo número extraño aparece por todas partes en matemáticas, no solo en fórmulas relacionadas con círculos y esferas, y de hecho lleva a aguas muy profundas. Incluso así, las fórmulas de la escuela son importantes y reflejan los orígenes de π en la geometría griega.

Allí, uno de los grandes problemas era la tarea no resuelta de cuadrar el círculo. Esta expresión se suele utilizar coloquialmente

[14] http://www.numberworld.org/misc_runs/pi-5t/details.html.

para indicar una manera errónea de abordar algo, como tratar de encajar una clavija cuadrada en un agujero redondo. Como muchas expresiones comunes extraídas de la ciencia, el significado de esta ha cambiado con el tiempo[15]. En los tiempos de Grecia, tratar de cuadrar el círculo era una idea del todo razonable. La diferencia entre las dos formas —recta o curva— es totalmente irrelevante: problemas similares tienen soluciones válidas[16]. Sin embargo, con el tiempo resultó que este problema concreto no puede resolverse utilizando los métodos especificados. La demostración es ingeniosa y técnica, pero su naturaleza general es comprensible.

En matemáticas, cuadrar el círculo significa construir un cuadrado cuya *área* sea la misma que la de un círculo dado, utilizando los métodos tradicionales de Euclides. En realidad la geometría griega permite otros métodos, de modo que un aspecto del problema es precisar qué métodos hay que utilizar. La imposibilidad de resolver el problema es entonces una afirmación sobre las limitaciones de tales métodos; no implica que no podamos calcular el área de un círculo. Simplemente tenemos que encontrar otro enfoque. La demostración de la imposibilidad explica por qué los geómetras griegos y sus sucesores no consiguieron encontrar una construcción del tipo requerido: no existe ninguna. Visto en retrospectiva, esto

[15] Lo que menos me gusta en este contexto es «salto cuántico». En el habla coloquial indica algún paso adelante gigantesco, o un cambio enorme, como el descubrimiento de América por los europeos. Sin embargo, en la teoría cuántica un salto cuántico es tan minúsculo que ningún instrumento conocido puede observarlo directamente, un cambio cuyo tamaño es aproximadamente 0,000 … 01 con unos cuarenta ceros.

[16] Encontrar una disección finita de un cuadrado en un círculo se denomina el problema de cuadrar el círculo de Tarski. Miklós Laczovich lo resolvió en 1990. Su método no es constructivo y hace uso del axioma de elección. El número de piezas requerido es enorme, del orden de 10^{50}.

explica por qué tuvieron que introducir métodos más esotéricos. De modo que la solución, pese a ser negativa, aclara lo que de otro modo sería un gran rompecabezas histórico. También evita que la gente pierda el tiempo en una búsqueda continua de una construcción que no existe, excepto para algunos espíritus contumaces que lamentablemente parecen incapaces de entender este mensaje, por muy detalladamente que se explique[17].

En los *Elementos* de Euclides los métodos tradicionales para construir figuras geométricas son versiones idealizadas de dos instrumentos matemáticos: la regla y el compás. Siendo pedantes, los compases, —por la misma razón que uno corta el papel con tijeras, y no con una tijera—; pero yo seguiré el habla común y evitaré el plural. Estos instrumentos se utilizan para «dibujar» diagramas en una hoja de papel ideal, el plano euclídeo.

Su forma determina lo que pueden dibujar. Un compás consta de dos varillas rígidas, articuladas. Una tiene una punta aguda, la otra contiene un lápiz fino. El instrumento se utiliza para dibujar un círculo, o una parte del mismo, con un centro y un radio concretos. Una regla es más sencilla: tiene un borde recto y se utiliza para trazar una línea recta. A diferencia de las reglas que uno compra en una papelería, las reglas de Euclides no tienen marcas, y esta es una restricción importante para el análisis matemático de lo que se puede crear con ellas.

---

[17] Las estrafalarias afirmaciones de los cuadradores del círculo y trisectores del ángulo se exploran en profundidad en Underwood Dudley, *A Budget of Trisections*, Springer, 1987, y *Mathematical Cranks*, Mathematical Association of America, 1992. El fenómeno no es nuevo: véase Augustus De Morgan, *A Budget of Paradoxes*, Longmans, 1872; reimpreso por Books For Libraries Press, 1915.

La regla y el compás del geómetra son idealizaciones en un sentido muy simple: se supone que trazan líneas infinitamente delgadas. Más aún, las líneas rectas son exactamente rectas y los círculos son absolutamente redondos. El papel es perfectamente plano y uniforme. El otro ingrediente clave de la geometría de Euclides es la noción de punto, otro ideal. Un punto es una mota en el papel, pero es una imposibilidad física: no tiene tamaño. «Un punto —decía Euclides, en la primera frase de los *Elementos* —, es lo que no tiene partes». Esto suena un poco como un átomo, o si usted está al corriente de la física moderna, una partícula subatómica, pero comparados con un punto geométrico, aquellos son gigantescos. Sin embargo, desde una perspectiva humana cotidiana un punto ideal de Euclides, un átomo y una mota de lápiz en una hoja de papel son suficientemente parecidos para los fines de la geometría.

Estos ideales no son alcanzables en el mundo real, por muy cuidadosamente que se construyan los instrumentos y se afile el lápiz, y por muy liso que se haga el papel. Pero el idealismo puede ser una virtud porque estos requisitos hacen las matemáticas mucho más sencillas. Por ejemplo, dos líneas de lápiz se cruzan en una pequeña región difusa con forma de paralelogramo, pero las líneas matemáticas se cortan en un solo punto. Las intuiciones que se ganan con círculos y líneas ideales pueden transferirse a menudo a las líneas y círculos reales e imperfectos. Así es como las matemáticas hacen su magia.

Dos puntos determinan una línea (recta), la única que pasa por ellos. Para construir la línea se coloca la regla ideal de modo que

pase por los dos puntos y se desliza a lo largo de ella el lápiz ideal. Dos puntos determinan también un círculo: se escoge uno como centro y se coloca allí la punta del compás; luego se ajusta de modo que la punta del lápiz caiga sobre el otro punto. Luego se gira el lápiz en un arco, manteniendo fijo el punto central. Dos líneas rectas determinan un único punto, en el que se cruzan, a menos que sean paralelas, en cuyo caso no se cruzan, pero eso abre una caja de Pandora de cuestiones lógicas. Una recta y un círculo determinan dos puntos, si se cruzan; determinan un solo punto si la recta es tangente al círculo; y nada en absoluto si el círculo es demasiado pequeño para tocar a la línea. Del mismo modo, dos círculos se encuentran en dos puntos, uno o ninguno.

La distancia es un concepto fundamental en el tratamiento moderno de la geometría euclídea. La distancia entre dos puntos cualesquiera se mide a lo largo de la recta que los une. Euclides se las arregló para hacer que su geometría funcionara sin un concepto de distancia explícito; encontró una manera de decir que dos segmentos de recta tienen la *misma* longitud sin definir la propia longitud. De hecho, esto es fácil: basta con extender un compás entre los extremos de un segmento, transferirlo al segundo segmento y ver si coincide. Si lo hace, las longitudes son iguales; si no, no lo son. En ningún momento se mide una longitud real.

Con estos ingredientes básicos los geómetras pueden construir formas y configuraciones más interesantes. Tres puntos determinan un triángulo a menos que estén sobre la misma recta. Cuando dos rectas se cortan, forman un ángulo. Un ángulo recto es

especialmente importante; una recta corresponde a dos ángulos rectos unidos. Y así sucesivamente, una y otra vez. Los *Elementos* de Euclides constan de trece libros, que ahondan cada vez más en las consecuencias de estos sencillos comienzos.

El grueso de los *Elementos* consiste en teoremas, propiedades válidas de la geometría. Pero Euclides también explica cómo resolver problemas geométricos, utilizando «construcciones» basadas en regla y compás. Dados dos puntos unidos por un segmento de recta, construir su punto medio. O trisecar el segmento: construir un punto a exactamente un tercio de camino a lo largo del mismo. Dado un ángulo, construir uno que lo biseca —tiene la mitad de tamaño—. Pero algunas construcciones sencillas se mostraron evasivas. Dado un ángulo, construir uno que lo triseca —tiene un tercio del tamaño—. Puede hacerse para segmentos de recta, pero nadie pudo encontrar un método para ángulos. Aproximaciones tan contiguas como se quiera, sí. Construcciones exactas utilizando solamente una regla sin marcas y un compás, no. Sin embargo, nadie necesita en realidad trisecar ángulos exactamente, de modo que esta cuestión particular no causaba muchos problemas.

Más complicada era una construcción que no podía ser ignorada: dado un círculo, construir un cuadrado que tenga la misma área. Este es el problema de cuadrar el círculo. Desde el punto de vista griego, si uno no pudiera resolverlo no estaría capacitado para afirmar que un círculo *tenía* un área. Y eso a pesar de que visiblemente encierra un espacio bien definido y la noción intuitiva de área es *cuánto espacio*. Euclides y sus sucesores, en particular

Arquímedes, convinieron una solución pragmática: suponer que los círculos tienen áreas, pero no esperar que seamos capaces de construir cuadrados con la misma área. Aun así, se pueden decir muchas cosas; por ejemplo, se puede demostrar, con todo rigor lógico, que el área de un círculo es proporcional al cuadrado de su diámetro. Lo que no se puede hacer, sin cuadrar el círculo, es construir una recta cuya longitud es la constante de proporcionalidad.

Los griegos no podían cuadrar el círculo utilizando regla y compás, de modo que establecieron otros métodos. Uno de ellos utilizaba una curva llamada una cuadratriz[18]. La importancia que daban a utilizar solo regla y compás fue exagerada por algunos comentaristas posteriores, y ni siquiera está claro que los griegos consideraran que cuadrar el círculo fuera una cuestión vital. Para el siglo XIX, sin embargo, el problema se estaba convirtiendo en una molestia importante. Unas matemáticas incapaces de responder a

[18] La cuadratriz de Hippias es la curva que traza el punto de intersección de una recta vertical que se mueve uniformemente a través de un rectángulo y una recta que rota uniformemente en torno al punto central del lado inferior del rectángulo (véase Figura 52). Esta relación convierte cualquier pregunta sobre división de ángulos en la correspondiente sobre división de rectas. Por ejemplo, para trisecar un ángulo simplemente se triseca la recta correspondiente. Véase: http://www.geom.uiuc.edu/~huberty/math5337/groupe/quadratrix.html

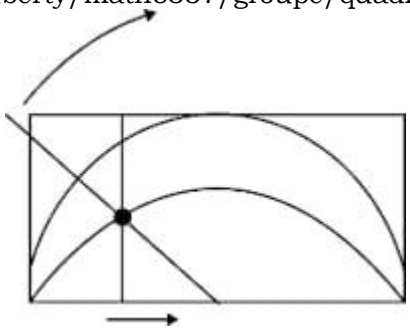


Figura 52. La cuadratriz de Hippias (curva inferior).

una cuestión tan simple eran como un cocinero con tres estrellas que no supiera cómo hervir un huevo.

Cuadrar el círculo suena a problema de geometría. Lo hace porque es un problema de geometría. Pero sucedió que su solución no estaba en la geometría en absoluto, sino en el álgebra. A veces, la clave para la solución de un gran problema está en establecer conexiones inesperadas entre áreas de las matemáticas en apariencia inconexas. Aquí, la conexión no carecía por completo de precedentes, pero su vínculo con la cuadratura del círculo no fue apreciado inicialmente. Incluso cuando lo fue, había una dificultad técnica, y tratarla requería aún otra área de las matemáticas: el análisis, la versión rigurosa del cálculo infinitesimal. Irónicamente, el primer avance importante vino de una cuarta área: la teoría de números. Y resolvía un problema geométrico para el que los griegos en el mejor de sus sueños nunca habían creído tener una solución, y en el que, hasta donde podemos decir, nunca pensaron: cómo construir, con regla y compás, un polígono regular con 17 lados.

Suena loco, sobre todo si añado que no existen construcciones semejantes para polígonos regulares con 7, 9, 11, 13 o 14 lados, pero sí existen para 3, 4, 5, 6, 8, 10 y 12. Sin embargo, hay un método detrás de la locura, y es el método que enriqueció las matemáticas.

Primero: ¿qué es un polígono regular? Un polígono es una figura acotada por líneas rectas. Es regular si dichas rectas tienen la misma longitud y se cortan a ángulos iguales. El ejemplo más familiar es el cuadrado: los cuatro lados tienen la misma longitud y

los cuatro ángulos son ángulos rectos. Hay otras formas con cuatro lados iguales o cuatro ángulos iguales: el rombo y el rectángulo, respectivamente. Solo un cuadrado tiene ambas propiedades. Un polígono regular de 3 lados es un triángulo equilátero, un polígono regular de 5 lados es un pentágono regular, y así sucesivamente (véase Figura 4). Euclides proporciona construcciones con regla y compás para polígonos regulares con 3, 4 y 5 lados. Los griegos también sabían cómo duplicar repetidamente el número de lados, lo que da 6, 8, 10, 12, 16, 20, y así sucesivamente. Combinando las construcciones para polígonos regulares de 3 y 5 lados pudieron obtener uno de 15 lados. Pero hasta aquí llegaba su conocimiento. Y así quedó durante unos dos mil años. Nadie imaginaba que fueran factibles otros números. Ni siquiera se lo preguntaban, simplemente parecía obvio que no podía hacerse nada más.

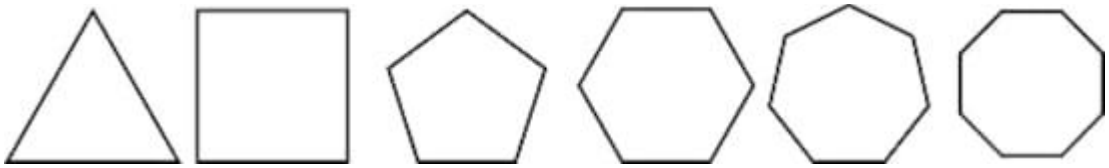


Figura 4. Los primeros polígonos regulares. De izquierda a derecha: triángulo equilátero, cuadrado, pentágono, hexágono, heptágono, octógono.

Se necesitó uno de los más grandes matemáticos que han existido nunca para pensar lo impensable, preguntar lo impreguntable y descubrir una respuesta verdaderamente sorprendente. A saber, Carl Friedrich Gauss. Gauss nació en una familia humilde de clase

obrera en la ciudad de Braunschweig (Brunswick) en Alemania. Su madre Dorotea no sabía leer ni escribir, y no pudo escribir la fecha de nacimiento de su hijo, pero recordaba que fue un miércoles, ocho días antes de la festividad de la Ascensión, en 1777. Más tarde Gauss calculó la fecha exacta a partir de una fórmula matemática que ideó para la fecha de la Pascua. Su padre Gebhard procedía de una familia de granjeros, pero se ganaba la vida con una serie de trabajos de bajo nivel: jardinero, peón, carnicero callejero, contable de funeraria. Su hijo fue un niño prodigio del que se dice que corregía la aritmética de su padre cuando tenía tres años, y sus aptitudes, que se extendían a las lenguas tanto como a las matemáticas, llevaron al duque de Braunschweig a financiar sus estudios universitarios en el Collegium Carolinum. Siendo todavía estudiante, Gauss redescubrió independientemente varios teoremas matemáticos importantes que habían sido demostrados por personas ilustres tales como Euler. Pero su teorema sobre el polígono regular de 17 lados llegó como un relámpago.

Para entonces hacía 140 años que se entendía el estrecho vínculo entre geometría y álgebra. En un apéndice al *Discours de la Méthode* («Discurso del método») René Descartes formalizó una idea que había estado flotando en forma rudimentaria durante algún tiempo: la idea de un sistema de coordenadas. En efecto, este toma el plano desnudo de Euclides, una hoja de papel en blanco, y lo convierte en un papel cuadriculado, lo que los ingenieros y los científicos llaman papel de gráficos. Se dibujan dos líneas rectas en el papel, una horizontal y otra vertical: se les llama ejes. Ahora se puede fijar la

localización de cualquier punto del plano preguntando a qué distancia está a lo largo del eje horizontal, y a qué distancia a lo largo del eje vertical (véase Figura 5, *izquierda*). Estos dos números, que pueden ser positivos o negativos, proporcionan una descripción completa del punto, y se denominan sus coordenadas.

Todas las propiedades geométricas de puntos, rectas, círculos y demás pueden traducirse en proposiciones algebraicas sobre las correspondientes coordenadas. Es muy difícil hablar significativamente sobre estas conexiones sin utilizar álgebra real —igual que es difícil hablar razonablemente sobre fútbol sin mencionar la palabra «gol»—. Así que algunas de las páginas que siguen incluirán algunas fórmulas.

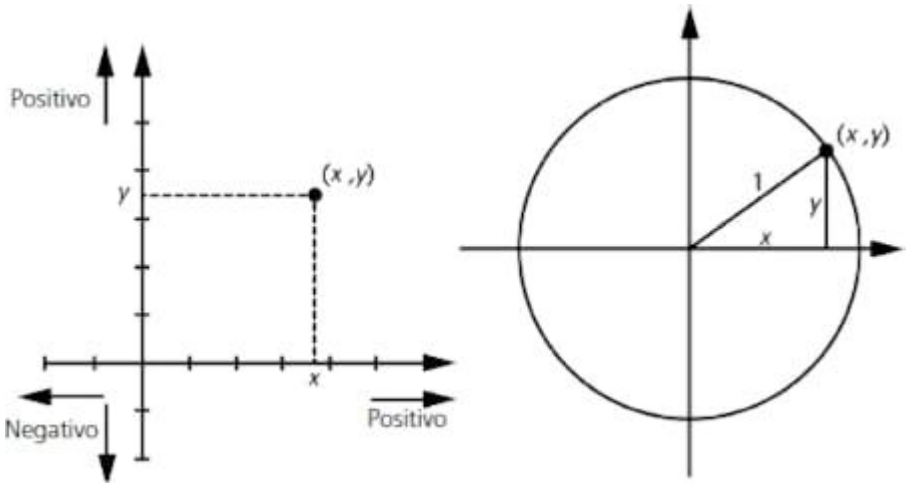


Figura 5. Izquierda: Coordenadas en el plano. Derecha: Cómo obtener la ecuación para el círculo unidad.

Están allí para asegurar que los protagonistas del drama tienen nombres y que la relación entre ellos está clara. «Romeo» es mucho más fácil de seguir que «el hijo de un patriarca italiano que se

enamora de la hermosa hija del enemigo jurado de su padre». Nuestro Romeo llevará el prosaico nombre *x*, y su Julieta será *y*.

A modo de ejemplo de cómo la geometría se convierte en álgebra, la Figura 5 (*derecha*) muestra cómo encontrar la ecuación de un círculo de radio unidad centrado en el origen, donde se cortan los dos ejes. El punto marcado tiene coordenadas (*x, y*), de modo que el triángulo rectángulo en la figura tiene un lado horizontal de longitud *x* y un lado vertical de longitud *y*. El lado más largo del triángulo es el radio del círculo, que es 1. El teorema de Pitágoras nos dice ahora que la suma de los cuadrados de las coordenadas es 1. En símbolos, un punto de coordenadas *x* e *y* se encuentra en el círculo si (y solo si) satisface la condición $x^2 + y^2 = 1$. Esta caracterización simbólica del círculo es breve y precisa, y muestra que realmente estamos hablando de álgebra. De modo recíproco, cualquier propiedad algebraica de pares de números, cualquier ecuación que incluye a *x* e *y*, puede reinterpretarse como una proposición geométrica sobre puntos, rectas, círculos o curvas más elaborados[19].

[19] He aquí un ejemplo explícito. Desde el punto de vista geométrico, si una línea corta a un círculo y no es tangente al mismo, entonces corta al círculo en exactamente dos puntos. Consideremos una línea que es paralela al eje horizontal, a distancia ½ por encima del mismo (véase Figura 53). La ecuación de esta línea es muy sencilla: $y = ½$. (Cualquiera que sea el valor de *x*, siempre obtenemos el mismo valor para *y*). Cuando $y = ½$, la ecuación $x^2 + y^2 = 1$ se hace $x^2 + ¼ = 1$. Por lo tanto $x^2 = 3/4$, de modo que $x = \sqrt{3}/2$ o $x = -\sqrt{3}/2$. Así, el álgebra nos dice que el círculo unidad corta a nuestra recta escogida exactamente en dos puntos, cuyas coordenadas son $(\sqrt{3}/2, ½)$ y $-(\sqrt{3}/2, ½)$. Esto es consistente con la Figura 53 y con un razonamiento puramente geométrico.

Las ecuaciones básicas del álgebra incluyen polinomios, combinaciones de potencias de una incógnita x, donde cada potencia está multiplicada por algún número, llamado coeficiente. La máxima potencia de x que aparece es el grado del polinomio. Por ejemplo, la ecuación

$$x^4 - 3x^3 - 3x^2 + 15x - 10 = 0$$

incluye un polinomio que empieza por x^4, de modo que su grado es 4. Los coeficientes son 1, -3, -3, 15 y -10. Hay cuatro soluciones distintas: $x = 1, 2, \sqrt{5}$ y $-\sqrt{5}$. Para estos valores el primer miembro de la ecuación es igual a cero, el segundo miembro. De los polinomios de grado 1, como $7x + 2$, se dice que son lineales, y solo incluyen la primera potencia de la incógnita. Las ecuaciones de grado 2, como $x^2 - 3x + 2$, se llaman ecuaciones cuadráticas, e incluyen la segunda potencia, el cuadrado. La ecuación para un círculo incluye una segunda variable, y. Sin embargo, si conocemos una segunda

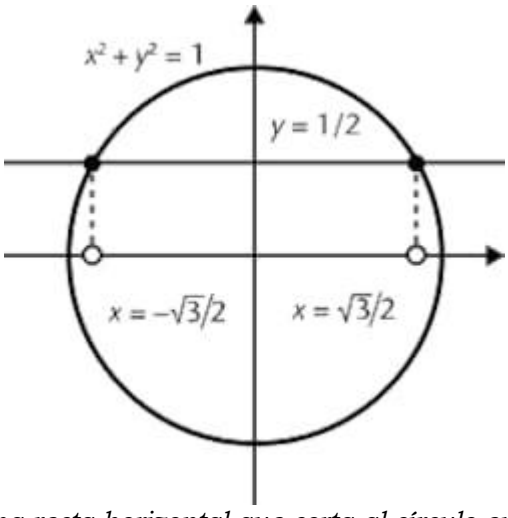


Figura 53. Una recta horizontal que corta al círculo en dos puntos.

ecuación que relaciona x e y, por ejemplo la ecuación que define una línea recta, entonces podemos escribir y en términos de x y reducir la ecuación de un círculo a una que solo incluye a x. Esta nueva ecuación nos dice dónde la recta corta al círculo. En este caso la nueva ecuación es cuadrática, con dos soluciones; así es como el álgebra refleja la geometría, en la que una recta corta al círculo en dos puntos distintos.

Esta propiedad del álgebra tiene una consecuencia importante para construcciones de regla y compás. Una construcción semejante, por complicada que sea, se descompone en una secuencia de pasos sencillos. Cada paso produce nuevos puntos en lugares donde se cortan dos rectas, dos círculos o una recta y un círculo. Estas líneas y círculos están determinados por puntos previamente construidos. Traduciendo la geometría en álgebra puede demostrarse que la ecuación algebraica que corresponde a la intersección de dos rectas es siempre lineal, mientras que en el caso de una recta y un círculo, o dos círculos, es cuadrática. En definitiva esto sucede porque la ecuación para un círculo incluye x^2 pero no potencias superiores de x. De modo que cada paso individual en una construcción corresponde a resolver una ecuación de grado 1 o grado 2 solamente.

Las construcciones más complejas son secuencias de estas operaciones básicas, y un poco de técnica algebraica nos permite deducir que cada coordenada de un punto que puede construirse por regla y compás es una solución de una ecuación polinómica, con coeficientes enteros, cuyo grado es una potencia de 2. Es decir,

el grado tiene que ser uno de los números 1, 2, 4, 8, 16, y así sucesivamente[20]. Esta condición es necesaria para que exista una construcción, pero puede reforzarse para dar una caracterización precisa de qué polígonos regulares son construibles. De repente, de una condición geométrica confusa y complicada emerge una rígida condición algebraica —que se aplica a cualquier construcción—. Ni siquiera se necesita saber cuál es la construcción: simplemente que solo utiliza regla y compás.

Gauss era consciente de esta idea elegante. También sabía (de hecho, cualquier matemático competente se habría dado cuenta de inmediato) que la cuestión de qué polígonos regulares pueden construirse mediante regla y compás se reduce a un caso especial cuando el polígono tiene un número primo de lados. Para ver por qué, pensemos en un número compuesto como 15, que es 3×5. Cualquier construcción hipotética de un polígono regular de 15 lados da automáticamente uno de 3 lados (nos quedamos con un vértice de cada cinco) y uno de 5 lados (nos quedamos con un vértice de cada 3), véase Figura 6. Con algún esfuerzo más se pueden combinar construcciones para un 3-gono y un 5-gono para obtener un 15-gono[21]. Los números 3 y 5 son primos, y la misma

[20] Estrictamente hablando, el polinomio en cuestión debe tener coeficientes enteros y ser irreducible: no es producto de dos polinomios de grado menor con coeficientes enteros. Tener un grado que es una potencia de 2 no siempre es suficiente para que exista una construcción con regla y compás, pero siempre es necesario. Si el grado no es una potencia de 2, no puede existir una construcción. Si es una potencia de 2, se necesita más análisis para decidir si existe una construcción.

[21] La recíproca también es cierta: dadas construcciones para 3-gonos y 5-gonos regulares, puede derivarse una para un 15-gono. La idea subyacente es que 2/5 - 1/3 = 1/15. Un punto sutil concierne a las potencias primas. El argumento no proporciona una construcción para, digamos, un 9-gono dada una para sus factores primos; a saber, un 3-gono. Gauss demostró que no es posible una construcción para potencias primas impares mayores que la primera.

idea se aplica en general. Así que Gauss se centró en polígonos con un número primo de lados, y se preguntó a qué se parecía la ecuación relevante. La respuesta era sorprendentemente limpia. Construir un polígono regular de 5 lados, por ejemplo, es equivalente a resolver la ecuación $x^5 - 1 = 0$. Reemplacemos 5 por cualquier otro primo y la proposición correspondiente es verdadera.

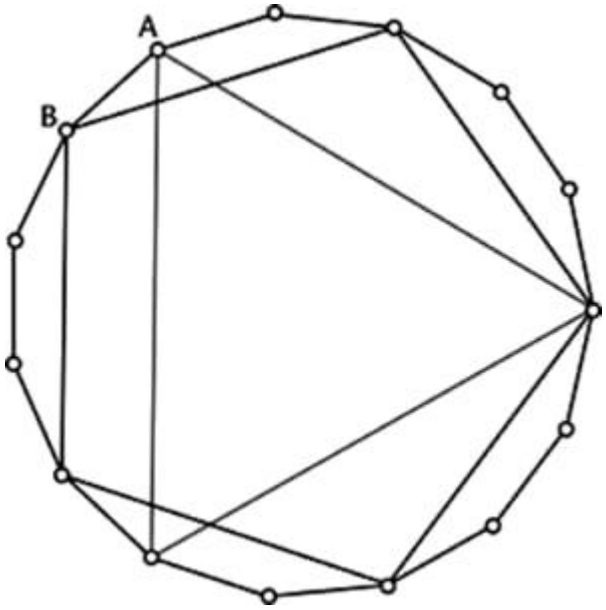


Figura 6. Construcción de un triángulo equilátero y un pentágono regular a partir de un 15-gono regular. Para la inversa, observemos que A y B son puntos consecutivos del 15-gono regular.

El grado de este polinomio es 5, que no es una de las potencias de 2 que yo he listado; incluso así, existe una construcción. Gauss descubrió rápidamente por qué: la ecuación se separa en dos partes, una de grado 1 y otra de grado 4. Tanto 1 como 4 son potencias de 2, y resulta que la ecuación de grado 4 es la crucial. Para ver por qué, tenemos que conectar la ecuación con la

geometría. Esto implica un nuevo tipo de número, uno que es generalmente olvidado en las matemáticas escolares pero que es indispensable para cualquier cosa que va más allá. Se llaman números complejos y su característica definitoria es que en el sistema de números complejos -1 tiene una raíz cuadrada.[22]

Un número «real» ordinario es positivo o negativo, y en uno u otro caso su cuadrado es positivo, de modo que -1 no puede ser el cuadrado de ningún número real. Esto es tan molesto que los matemáticos inventaron un tipo nuevo de número «imaginario» cuyo cuadrado es -1. Necesitaban un nuevo símbolo para el mismo, así que lo llamaron «i» (por «imaginario»). Las operaciones habituales del álgebra —sumar, restar, multiplicar, dividir— llevan a combinaciones de números reales e imaginarios tales como 3 + 2i. Se dice que son complejos, lo que no quiere decir «complicados» sino que indica que tienen dos partes: 3 y 2i. Los números reales se encuentran sobre la famosa recta real, como los números en una regla. Los números complejos yacen en un plano de números, en el que se coloca una regla imaginaria perpendicular a una real, y las dos forman un sistema de coordenadas (véase Figura 7, *izquierda*).

Durante los últimos doscientos años los matemáticos han considerado que los números complejos son fundamentales para su disciplina. Ahora reconocemos que lógicamente están en pie de igualdad con los más familiares números «reales» —que, como todas las estructuras matemáticas, son conceptos abstractos, no objetos físicos reales—. Los números complejos se utilizaban ampliamente

[22] Véase Ian Stewart, *Seventeen Equations that Changed the World*, Profile, 2012, capítulo 5.

antes de Gauss, pero su estatus era todavía misterioso hasta que Gauss y algunos otros los desmitificaron. La fuente de su atractivo era paradójica: pese al misterio que rodeaba a su significado, los números complejos se comportaban mucho mejor que los números reales. Suministraban un ingrediente del que carecían los números reales. Proporcionaban un conjunto completo de soluciones para una ecuación algebraica.

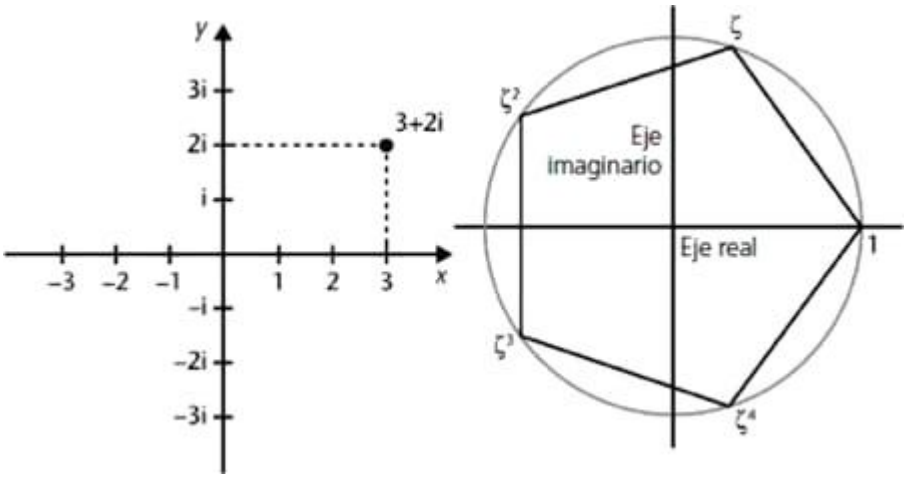


Figura 7. Izquierda: El plano complejo. Derecha: Las raíces quintas complejas de la unidad.

Las ecuaciones cuadráticas son el ejemplo más sencillo. Algunas cuadráticas tienen dos soluciones reales, mientras que otras no tienen ninguna. Por ejemplo $x^2 - 1 = 0$ tiene las soluciones 1 y -1, pero $x^2 + 1 = 0$ no tiene soluciones. En medio está $x^2 = 0$, cuya única solución es 0, pero en cierto sentido esta es la misma solución «repetida dos veces[23]». Sin embargo, si admitimos soluciones

[23] Para dar sentido a esta afirmación, resolvemos la cuadrática en factores lineales. Entonces $x^2 - 1 = (x + 1)(x - 1)$ que es cero si uno de los factores es cero, de modo que $x = 1$ o -1. El mismo

complejas, entonces $x^2 + 1 = 0$ tiene también dos soluciones: i y -i. Gauss no tenía reparos en utilizar números complejos; de hecho, su tesis doctoral proporcionó la primera demostración lógicamente correcta del teorema fundamental del álgebra: el número de soluciones complejas para cualquier ecuación polinómica (con multiplicidades contadas correctamente) es igual al grado de la ecuación. De modo que las cuadráticas (grado 2) siempre tienen dos soluciones complejas, las cúbicas (grado 3) siempre tienen tres soluciones complejas, y así sucesivamente.

La ecuación $x^5 - 1 = 0$, que afirmé que define un pentágono regular, tiene grado 5. Por consiguiente tiene cinco soluciones complejas. Hay solamente una solución real: $x = 1$. ¿Qué pasa con las otras cuatro? Proporcionan cuatro vértices de un pentágono regular perfecto en el plano complejo, siendo $x = 1$ la quinta (véase Figura 7, *derecha*). Esta correspondencia es un ejemplo de belleza matemática: una forma geométrica elegante se convierte en una ecuación elegante.

Ahora bien, la ecuación cuyas soluciones son estos cinco puntos tiene grado 5, que no es una potencia de 2. Pero, como se ha mencionado antes, la ecuación de grado 5 se divide en dos partes con grados 1 y 4, llamadas sus factores irreducibles:

$$x^5 - 1 = (x - 1)(x^4 + x^3 + x^2 + x + 1)$$

---

razonamiento puede aplicarse a $x^2 = xx$: esto es cero si o bien el primer factor $x = 0$ o bien el segundo $x = 0$. Sucede así que estas dos soluciones dan el mismo x, pero la presencia de dos factores x distingue esta situación de algo como $x(x - 1)$ donde solo hay un factor x. Al contar cuantas soluciones tiene una ecuación algebraica, la respuesta es generalmente mucho más rigurosa si se tienen en cuenta estas «multiplicidades».

(«Irreducible» significa que no existen más factores, como sucede con los números primos). El primer factor da la solución real $x = 1$. El otro factor da las cuatro soluciones complejas y los otros cuatro vértices del pentágono. De modo que todo cobra mucho más sentido, y es mucho más elegante, cuando utilizamos números complejos.

Suele ser difícil reconstruir cómo llegaron los matemáticos del pasado a nuevos descubrimientos, porque ellos tenían la costumbre de presentar solamente el resultado final de sus deliberaciones y no los muchos pasos en falso que dieron en el camino. Este problema suele ser complicado, porque las pautas de pensamiento naturales en épocas pasadas eran diferentes de las actuales. Gauss en particular es lamentablemente famoso por esconder sus huellas y publicar solo su análisis final y muy depurado. Pero cuando se trata de la investigación de Gauss sobre el polígono de 17 lados estamos en terreno seguro; el análisis final que publicó proporciona varias claves útiles.

Su punto de partida no era nuevo. Varios matemáticos anteriores eran perfectamente conscientes de que el análisis anterior de los pentágonos regulares funcionaba con toda generalidad. Construir un polígono con cualquier número n de lados es equivalente a resolver la ecuación $x^n - 1 = 0$ en números complejos. Además, este polinomio factoriza como

$$(x - 1)(x^{n-1} + x^{n-2} + \ldots + x^2 + x + 1)$$

De nuevo el primer factor da la solución real $x = 1$ y las $n - 1$ soluciones restantes proceden del segundo factor. Cuando n es impar, estas son todas complejas; cuando n es par, una de ellas es una segunda solución real $x = -1$.

Lo que Gauss advirtió, y todos los demás habían pasado por alto, es que a veces el segundo factor puede expresarse utilizando una serie de ecuaciones cuadráticas. No representándolo como un producto de factores más simples, porque eso no es posible, sino utilizando ecuaciones cuyos coeficientes resuelven otras ecuaciones. El hecho clave aquí —el punto débil del problema— es una propiedad elegante de las ecuaciones algebraicas, que surge cuando resolvemos varias de ellas de una en una de esta manera. El cálculo es siempre equivalente a resolver una sola ecuación, pero en general el grado se hace mayor. Por lo tanto, el precio que pagamos por tener menos ecuaciones es un incremento en el grado. Puede ser confuso, pero hay una propiedad que podemos predecir: cuán grande se hace el grado. Simplemente se multiplican los grados de los polinomios sucesivos.

Si todos son cuadráticos, el resultado es $2\times2\times\ldots\times2$, una potencia de 2. De modo que $n - 1$ debe ser una potencia de 2 si existe una construcción. Sin embargo, esta condición no siempre es suficiente. Cuando $n = 9$, $n - 1 = 8$, que es una potencia de 2. Pero Gauss descubrió que no existe ninguna construcción para el 9-gono

regular. La razón es que 9 no es primo[24]. ¿Qué pasa con el caso siguiente, en el que resolvemos una serie de cuatro ecuaciones cuadráticas? Ahora el grado n - 1 de la única ecuación correspondiente es 2×2×2×2 = 16. De modo que n = 17, y este sí es primo.

En este momento Gauss debe haber sabido que estaba sobre algo, pero hay otro punto técnico, posiblemente fatal. Gauss se había convencido de que para que exista la construcción de un pentágono regular con un número primo de lados, ese primo debe ser una potencia de 2, más 1. De modo que esta condición es necesaria para que exista una construcción: si falla, no hay tal construcción. Sin embargo, la condición podría no ser suficiente: de hecho hay muchas ecuaciones de grado 16 que no se reducen a una serie de cuatro cuadráticas.

No obstante, había una razón para ser optimista: las construcciones griegas. ¿Qué primos aparecían allí? Solo tres: 2, 3 y 5. Todos son una potencia de 2 más 1, a saber, $2^0 + 1$, $2^1 + 1$ y $2^2 + 1$. El álgebra asociada con el pentágono proporciona otras claves. Pensando en esto, Gauss demostró que el polinomio de grado 16 asociado con el polígono de 17 lados puede reducirse a una serie de polinomios cuadráticos. Por consiguiente debe existir una construcción mediante regla y compás. Un método similar demostraba que lo mismo es cierto cuando el número de lados es un primo que es alguna potencia de 2 más 1. Las ideas son un tributo a la capacidad

[24] Cuando n = 9, el segundo factor es $x^8 + x^7 + x^6 + x^5 + x^4 + x^3 +$ [illegible] $+ x + 1$ Pero esto tiene factores: es igual a $(x^2 + x + 1)$ ($x^6 +$ [illegible] $+ 1$). La caracterización de Gauss de números construibles requiere que cada factor irreducible tenga un grado que sea una potencia de 2. Pero el segundo factor tiene grado 6, que no es una potencia de 2.

de Gauss para entender pautas matemáticas. En su núcleo hay algunos teoremas generales de la teoría de números, en los que no voy a entrar. Lo importante es que nada de esto era accidental. Había sólidas razones estructurales para que funcionara. Solo había que ser un Gauss para advertirlas.

Gauss no proporcionó una construcción explícita, pero dio una fórmula para las soluciones de la ecuación de grado 16 que puede transformarse en una construcción semejante si realmente se quiere una[25]. Cuando desarrolló sus ideas en las *Disquisitiones*

[25] Gauss demostró que el 17-gono puede construirse siempre que se pueda construir una recta cuya longitud sea

$$\frac{1}{16}\left[-1+\sqrt{17}+\sqrt{34-2\sqrt{17}}\right.$$

$$\left.+\sqrt{68+12\sqrt{17}-16\sqrt{34+2\sqrt{17}}-2\left(1-\sqrt{17}\right)\sqrt{34-2\sqrt{17}}}\right]$$

Puesto que siempre se pueden construir raíces cuadradas, esto resuelve efectivamente el problema. Otros matemáticos encontraron construcciones explícitas. Ulrico von Huguenin publicó la primera en 1803, y H. W. Richmond encontró una más sencilla en 1893. En la Figura 54, tomar dos radios perpendiculares AOP $_0$ y BOC de un círculo. Hacer OJ = 1/4OB y ángulo OJE = 1/4OJP $_0$. Encontrar F tal que el ángulo EJF es 45 °. Trazar un círculo con FP $_0$ como diámetro, que corta a OB en K. Trazar el círculo con centro E que pasa por K, que corta a AP $_0$ en G y H. Trazar HP $_3$ y GP $_5$ perpendicular a AP $_0$. Entonces, P $_0$, P $_3$, P $_5$ son respectivamente los vértices 0-ésimo, tercero y quinto de un 17-gono regular, y los otros vértices se construyen ahora fácilmente.

arithmeticae omitió algunos detalles, pero afirmó que poseía demostraciones completas. Su descubrimiento épico le convenció de que debía dedicar su vida a las matemáticas antes que a las lenguas. El duque continuó apoyando financieramente a Gauss, pero Gauss quería algo más permanente y fiable. Cuando el astrónomo Giuseppe Piazzi descubrió el primer asteroide, Ceres, solo pudieron hacerse unas pocas observaciones antes de que este nuevo mundo se hiciera invisible contra el brillo del Sol. A los astrónomos les preocupaba que no fueran capaces de encontrarlo de nuevo. En un *tour de force* que incluía nuevas técnicas para calcular órbitas, Gauss predijo dónde iba a reaparecer… y tenía razón. Esto le llevó a ser nombrado profesor de astronomía y director del Observatorio de Gotinga. Continuó ocupando este puesto durante el resto de su vida.

El caso es que 17 no es el único nuevo número de este tipo. Se conocen dos más: $2^8 + 1 = 257$ y $2^{16} + 1 = 65.537$. (Un poco de álgebra muestra que la potencia de 2 que ocurre debe ser ella

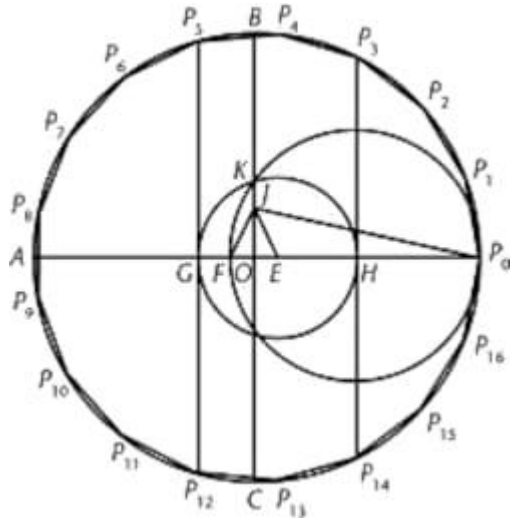


Figura 54. Cómo construir un 17-gono regular.

misma una potencia de 2; si no lo es, entonces el número no puede ser primo). Sin embargo, la pauta se detiene en este punto, porque $2^{32} + 1 = 4.294.967.297$ es igual a 641×6.700.417, luego no es primo. Se sabe que los denominados números de Fermat $2^{2^n} + 1$ no son primos para $n = 5, 6, 7...$, hasta 32. También se sabe que muchos números de Fermat más grandes no son primos. No se han encontrado otros números de Fermat primos, pero su existencia no es en absoluto imposible[26]. Se conoce una construcción para el polígono de 257 lados. Un matemático dedicó muchos años al polígono de 65.537 lados, una tarea algo absurda, y en cualquier caso sus resultados contienen errores[27].

El resultado del análisis de Gauss es que se puede construir un polígono regular con regla y compás si y solo si el número de lados es un producto de una potencia de 2 y *distintos* números primos de Fermat impares. En particular, un polígono regular de 9 lados no puede construirse de esta manera. Esto implica inmediatamente que al menos un ángulo no puede trisecarse, porque el ángulo en un triángulo equilátero tiene 60 °, y un tercio de esto es 20 °. Dado este ángulo, es fácil construir un polígono regular de 9 lados. Puesto que eso es imposible, no hay una construcción general de regla y compás para trisecar un ángulo.

Gauss omitió muchos detalles de las demostraciones cuando escribió sus resultados, y los matemáticos no podían aceptar su

[26] Para los últimos descubrimientos, véase Wilfrid Keller, «Factores primos de números de Fermat y estatus completo de factorización» http://www.prothsearch.net/fermat.html.

[27] E. J. Richelot publicó una construcción para el 257-gono regular en 1832. J. Hermes de la Lingen University dedicó diez años al polígono de 65.537 lados. Su trabajo no publicado puede encontrarse en la Universidad de Gotinga, pero se cree que contiene errores.

palabra sin más. En 1837 el matemático francés Pierre Wantzel publicó una demostración completa de la caracterización de Gauss de polígonos regulares construibles, y dedujo la imposibilidad de trisecar un ángulo general mediante construcción por regla y compás. También demostró que es imposible construir un cubo cuyo volumen es el doble del de un cubo dado, otro antiguo problema griego conocido como «la duplicación del cubo».

Tanto la trisección del ángulo como la duplicación del cubo resultan ser imposibles porque las longitudes implicadas satisfacen ecuaciones cúbicas irreducibles —de grado 3—. Puesto que 3 no es una potencia de 2, esto excluye directamente la cuestión. Sin embargo, este método no parecía funcionar para el problema de cuadrar el círculo, y ello por razones interesantes. Un círculo de radio unidad tiene área π, y un cuadrado de dicha área tiene lado √π. Existen construcciones geométricas para raíces cuadradas, y también construcciones para cuadrados, de modo que cuadrar el círculo se reduce a partir de una recta de longitud 1 y construir una de longitud π. Si resultara que π satisface una ecuación cúbica irreducible —o cualquier ecuación irreducible cuyo grado no es una potencia de 2— entonces los métodos de Wantzel probarían que es imposible cuadrar el círculo.

Sin embargo, nadie conocía ninguna ecuación algebraica que fuera satisfecha exactamente por π, y mucho menos una cuyo grado no es una potencia de 2. El valor de la escuela 22/7 satisface $7x - 22 = 0$, pero eso es solo una aproximación a π, ligeramente mayor, de modo que no sirve. Si pudiera probarse que no existe tal ecuación —y

muchos lo sospechaban basados en que si existiera habría sido encontrada— se seguiría de ello la imposibilidad de cuadrar el círculo. Por desgracia, nadie pudo demostrar que no hay tal ecuación. El estatus algebraico de π estaba en el limbo. La solución final utilizaba métodos que no solo iban más allá de la geometría, sino que también iban más allá del álgebra.

Para apreciar cuál es aquí la cuestión más importante tenemos que empezar con una idea más simple. Hay una distinción importante en matemáticas entre números que pueden expresarse como fracciones exactas p/q, donde p y q son números enteros, y los que no pueden expresarse de esa manera. De los primeros se dice que son racionales (son razones de números enteros), y los segundos son irracionales. Por ejemplo, la aproximación 22/7 a π es racional. Hay aproximaciones mejores; una famosa es 355/113, correcta hasta seis cifras decimales. Sin embargo, se sabe que ninguna fracción puede representar a π exactamente: es irracional. Esta propiedad sospechada durante largo tiempo fue demostrada por primera vez por el matemático suizo Johann Heinrich Lambert en 1768. Su demostración se basa en una ingeniosa fórmula para la función tangente en trigonometría, que él expresaba como una fracción continua: una pila infinita de fracciones ordinarias[28]. En 1873 Charles Hermite encontró una demostración más simple,

[28] Una típica fracción continua tiene este aspecto:

$$3+\cfrac{1}{7+\cfrac{1}{15+\cfrac{1}{1+\cfrac{1}{192+\cdots}}}}$$

Esta fracción continua concreta es el comienzo de la que representa π.

basada en fórmulas del cálculo infinitesimal, que iba más allá: demostraba que π^2 es irracional. Por consiguiente π tampoco es la raíz cuadrada de un número racional.

Lambert sospechaba algo mucho más fuerte. En el artículo en que demostraba que π es irracional, conjeturaba que π es trascendente; es decir, π no satisface ninguna ecuación polinómica con coeficientes enteros. Trasciende la expresión algebraica. Descubrimientos posteriores probaron que tenía razón. El avance fundamental llegó en dos etapas. El nuevo método de Hermite para demostrar la irracionalidad fijaba el escenario al sugerir que el cálculo infinitesimal —más exactamente, su versión rigurosa, el análisis— podría ser una estrategia útil. Desarrollando esta idea, Hermite encontró una demostración maravillosa de que el otro famoso y curioso número en matemáticas, la base e de los logaritmos naturales, es trascendente. El valor numérico de e es aproximadamente 2,71828, y si acaso es incluso más importante que π. La demostración de trascendencia de Hermite es magia, un conejo extraído con una floritura de la chistera del análisis. El conejo es una fórmula complicada relacionada con una hipotética ecuación algebraica que se supone que es satisfecha por e. Utilizando álgebra, Hermite demuestra que esta fórmula es igual a algún entero no nulo. Utilizando análisis, él demuestra que debe estar entre -1/2 y ½. Puesto que el único entero en este rango es cero, estos resultados son contradictorios. Por consiguiente, la hipótesis de que e satisface una ecuación algebraica debe ser falsa, luego e es trascendente.

En 1882 Ferdinad Lindemann añadió algunos elementos al método de Hermite, y demostró que si un número distinto de cero satisface una ecuación algebraica, entonces e elevado a la potencia de ese número no satisface una ecuación algebraica. Luego aprovechó una relación que era conocida para Euler que incluye a π, e y el número imaginario i: la famosa fórmula $e^{i\pi} = -1$. Supongamos que π satisface alguna ecuación algebraica. Entonces también lo hace iπ, y el teorema de Lindemann implica que -1 no satisface una ecuación algebraica. Sin embargo, es evidente que lo hace: es la solución de $x + 1 = 0$. La única salida a esta contradicción lógica es que π no satisfaga una ecuación algebraica; es decir, es trascendente. Y eso significa que no se puede cuadrar el círculo.

Fue un viaje largo e indirecto desde la geometría de Euclides hasta la demostración de Lindemann, y necesitó más de doscientos años, pero los matemáticos finalmente llegaron allí. La historia no solo nos dice que no se puede cuadrar el círculo. Es una lección de cómo llegan a resolverse grandes problemas matemáticos. Se requería que los matemáticos formularan cuidadosamente lo que entendían por «construcción geométrica». Tenían que fijar propiedades generales de tales construcciones que pudieran poner límites a lo que podían conseguir. Encontrar esas propiedades requería establecer conexiones con otra área de las matemáticas: el álgebra. Resolver el problema algebraico, incluso en casos más simples tales como la construcción de polígonos regulares, también implicaba la teoría de números. Tratar con el caso difícil de π requirió más innovaciones, y

el problema aún tenía que ser trasladado a otra área de las matemáticas: el análisis.

Ninguno de estos pasos era simple u obvio. Se necesitó casi un siglo para completar la demostración, incluso cuándo se tenían las ideas principales. Los matemáticos involucrados eran de los mejores de su tiempo, y al menos uno era de los mejores de todos los tiempos. Resolver grandes problemas requiere un conocimiento profundo de las matemáticas, más constancia e ingenio. Puede exigir años de esfuerzo concentrado, buena parte del cual es aparentemente infructuoso. Pero imaginemos lo que debe sentirse cuando la constancia da fruto y se abre una grieta en algo que ha desconcertado al resto de la humanidad durante siglos. Como dijo el presidente John F. Kennedy en 1962 cuando anunció el proyecto de ir a la Luna: «Nosotros decidimos… hacer [estas]… cosas, no porque sean fáciles, sino porque son difíciles».

Pocas historias tienen fin en matemáticas, y π no es una excepción. De vez en cuando aparecen nuevos descubrimientos asombrosos sobre π. En 1997 Fabrice Bellard anunció que la cifra decimal que ocupa el lugar un billón de π, en notación binaria, es 1.[29] Lo que hacía notable este resultado no era la respuesta. La propiedad asombrosa era que él no calculó ninguno de los dígitos anteriores. Simplemente sacó del aire un dígito concreto.

Lo que hizo posible el cálculo era una curiosa fórmula para π descubierta por David Bailey, Peter Borwein y Simon Plouffe en

[29] http://bellar4d.org/pi-challenge/announce220997.html.

1996. Puede parecer un poco complicada, pero echémosle una ojeada de todas formas:

$$\pi = \sum_{n=0}^{\infty} \frac{1}{2^{4n}} \left(\frac{4}{4n+1} - \frac{2}{4n+4} - \frac{1}{4n+5} - \frac{41}{4n+6} \right)$$

La ∑ grande significa «suma» sobre el rango especificado. Aquí *n* va de 0 a infinito (∞). Bellard utilizó en realidad una fórmula que había derivado utilizando métodos similares, que es algo más rápida para hacer cálculos.

$$\pi = \frac{1}{64} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{10n}} \left(-\frac{32}{4n+1} - \frac{1}{4n+3} + \frac{256}{10n+1} - \frac{64}{10n+3} - \frac{4}{10n+5} - \frac{4}{10n+7} + \frac{1}{10n+9} \right)$$

El punto clave es que muchos de los números que aparecen aquí —1, 4, 32, 64, 256 y también 2^{4n} y 2^{10n}— son potencias de 2, que por supuesto son muy simples en el sistema binario que se utiliza en los ordenadores. Este descubrimiento estimuló un diluvio de nuevas fórmulas para π, y para varios otros números interesantes. El récord para encontrar un simple dígito binario de π se bate regularmente: en 2010 Nicholas Sze de Yahoo calculó el dígito binario número dos trillones de π, que resulta ser 0.

Las mismas fórmulas pueden utilizarse para encontrar dígitos aislados de π en aritméticas con base 4, 8 y 16. Nada de este tipo se

conoce para ninguna otra base; en particular, no podemos computar dígitos decimales aislados. ¿Existen tales fórmulas? Hasta que se encontró la fórmula de Bailey-Borwein-Plouffe, nadie imaginaba que pudiera hacerse en sistema binario.

Capítulo 4

Cartografiando misterios

El teorema de los cuatro colores

Muchos de los grandes problemas matemáticos derivan de preguntas difíciles y profundas en áreas de la disciplina bien establecidas. Son los grandes desafíos que surgen cuando un área importante ha sido explorada exhaustivamente. Suelen ser bastante técnicos, y todos los que trabajan en el área saben que son difíciles de responder porque muchos expertos lo han intentado y han fracasado. El área en cuestión ya poseerá muchas técnicas poderosas, grandes máquinas matemáticas cuyas manivelas pueden manejarse si uno ha hecho su tarea. Pero si el problema sigue abierto es que todas las maneras plausibles de utilizar dichas técnicas ya han sido ensayadas y *no han funcionado*. Así que o bien hay una manera menos evidente de utilizar las técnicas del área ensayadas-y-comprobadas, o es necesario utilizar nuevas técnicas.
Se han dado ambos casos.

Otros grandes problemas son muy diferentes. Aparecen de la nada: unas rayas en la arena, un garabato en un margen, un capricho pasajero... Sus enunciados son simples, pero debido a que no tienen muchos antecedentes matemáticos, no hay métodos establecidos para pensar en ellos. Pueden pasar muchos años antes de que su dificultad se haga evidente: alguien podría descubrir un truco ingenioso aunque simple que los resolviera en media página. El problema de los cuatro colores es de este segundo tipo. Pasaron

décadas antes de que los matemáticos empezaran a entender cuán difícil era la pregunta, y durante una buena parte de este tiempo pensaban que *había* sido resuelto en unas pocas páginas. Parecía ser una cuestión marginal, y por ello muy pocos se molestaron en tomarla en serio. Cuando lo hicieron, la presunta solución resultó ser errónea. La solución final corrigió los fallos, pero para entonces el argumento se había hecho tan complicado que fue necesario recurrir a la ayuda intensiva de un ordenador.

A largo plazo, ambos tipos de problemas convergen, pese a sus diferentes fundamentos, porque resolverlos requiere nuevas maneras de pensar. Los problemas del primer tipo pueden estar encuadrados en un área bien entendida, pero los métodos tradicionales en dicha área son inadecuados. Los problemas del segundo tipo no pertenecen a ninguna área establecida —de hecho, motivan la creación de áreas nuevas— de modo que no hay métodos tradicionales que puedan utilizarse. En ambos casos, resolver el problema exige inventar nuevos métodos y forjar nuevos vínculos con el corpus de matemáticas existente.

Sabemos exactamente cuál fue el origen del problema de los cuatro colores, y no estaba en las matemáticas. En 1852 Francis Guthrie, un joven matemático y botánico surafricano que se preparaba para conseguir un grado en derecho, estaba intentando colorear los condados en un mapa de Inglaterra. Quería estar seguro de que a dos condados adyacentes cualesquiera se les asignaran colores distintos para que las fronteras se vieran con nitidez. Guthrie descubrió que solo necesitaba cuatro colores para completar la

tarea, y después de experimentar algún tiempo se convenció de que esto sería cierto para cualquier mapa. Por «adyacentes» él entendía que los condados concernidos compartían una frontera de longitud no nula: si dos países se tocaban en un punto, o en varios puntos aislados, podían tener el mismo color si fuera necesario. Sin esta salvedad, no hay límite para el número de colores, porque cualquier número de regiones pueden tocarse en un punto (véase Figura 8, *izquierda*).

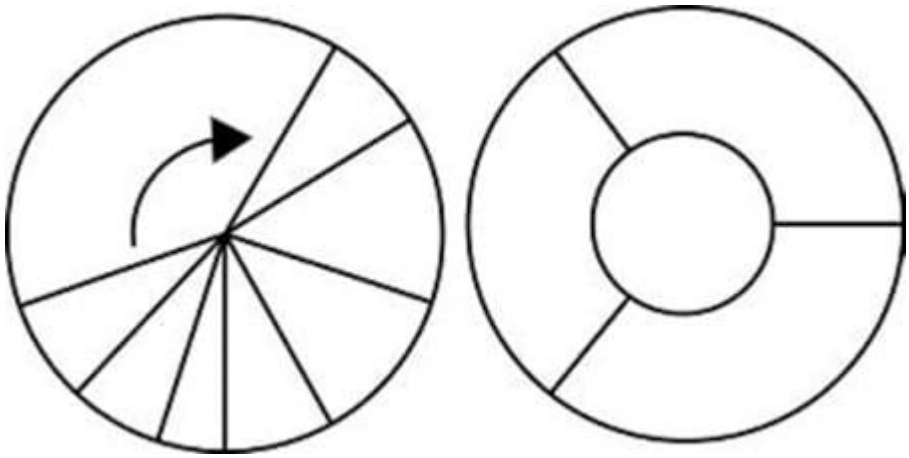


Figura 8. Izquierda: Cualquier número de regiones pueden tocarse en un punto. Derecha: Son necesarios al menos cuatro colores.

Preguntándose si esta proposición era un teorema matemático conocido, se lo planteó a su hermano Frederick, quien estaba estudiando matemáticas con el distinguido pero excéntrico Augustus de Morgan en el University College en Londres. De Morgan no lo sabía, así que escribió a un matemático aún más distinguido, el irlandés *sir* William Rowan Hamilton:

Uno de mis estudiantes [más tarde identificado como Frederick Guthrie] me ha pedido hoy que le diera una razón para un hecho que yo no sabía que era un hecho —y sigo sin saberlo—. Él dice que si se divide una figura y los compartimentos se colorean con diferentes colores de modo que figuras con una porción de frontera común tengan colores diferentes, pueden necesitarse cuatro colores pero no más… La pregunta es si puede inventarse algo que necesite cinco o más… ¿Qué dice usted? Y, si es cierto, ¿ha sido advertido?

Frederick se refirió más tarde a una «demostración» que había sugerido su hermano, pero también dijo que la idea clave era un dibujo equivalente a la Figura 8, que solo prueba que no bastará con menos de cuatro colores.

La respuesta de Hamilton fue breve y de poca ayuda. «Es poco probable que pueda intentar muy pronto su "cuaternión" de colores», escribió. En esa época él estaba trabajando en un sistema de álgebra que se convirtió en una obsesión para toda la vida; un sistema análogo a los números complejos pero que incluía cuatro tipos de números en lugar de los dos (reales e imaginarios) de los números complejos. Los llamó «cuaterniones». El sistema sigue siendo importante en matemáticas; de hecho, es sin duda más importante ahora que lo fue en tiempos de Hamilton. Pero nunca ha alcanzado realmente las alturas que esperaba Hamilton. Hamilton solo estaba haciendo una broma académica cuando utilizó la

palabra, y durante mucho tiempo no pareció haber un vínculo entre los cuaterniones y el problema de los cuatro colores. Sin embargo, hay una reformulación del problema que puede verse como una proposición sobre cuaterniones, de modo que la broma de Hamilton tiene un aguijón en la cola[30].

Siendo incapaz de encontrar una demostración, De Morgan mencionó el problema a sus amigos matemáticos con la esperanza de que alguno de ellos pudiera dar con una idea. A finales de la década de 1860, el lógico, matemático y filósofo norteamericano Charles Sanders Peirce afirmó que había resuelto el problema de los cuatro colores, junto con otras cuestiones similares sobre mapas en superficies más complicadas. Su presunta demostración nunca fue publicada, y es dudoso que los métodos de los que disponía hubieran sido adecuados.

Aunque el problema de los cuatro colores trata ostensiblemente de mapas, no tiene aplicaciones útiles en cartografía. Los criterios prácticos que se utilizan para colorear mapas reflejan sobre todo diferencias políticas, y si eso significa que regiones adyacentes deben tener el mismo color, así sea. El interés del problema está por completo dentro de las matemáticas puras, en una nueva área que empezaba a desarrollarse: la topología. Esta es «geometría-de-la-lámina-elástica» en la que las figuras pueden deformarse de cualquier manera continua. Pero incluso aquí, el problema de los

[30] Louis H. Kauffman, «Map coloring end the vector cross product, *Journal of Combinatorial Theory*», B 48, (1990), pp. 145-154. Louis H. Kauffman, «Reformulating the map color theoreme», *Discrete Mathematics*, n.º 302 (2005), pp. 145-172.

cuatro colores no pertenecía a la corriente principal. Parecía ser tan solo una curiosidad menor.

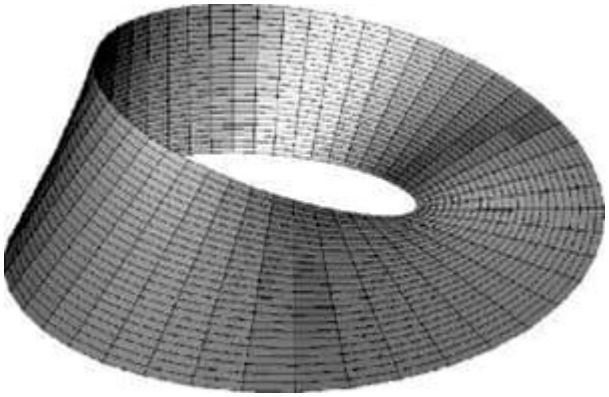


Figura 9. La banda de Möbius tiene solo un lado.

Uno de los pioneros de la topología fue August Möbius, hoy famoso por su banda de un lado (véase Figura 9). Es fácil hacer un modelo tomando una cinta de papel, curvándola en forma de anillo como un cilindro corto y grueso, girando 180 ° un extremo y pegándolo al otro. Un amigo de Möbius, el lingüista Benjamin Weiske, le planteó un rompecabezas: ¿podía un rey indio con cinco hijos, todos ellos príncipes, dividir su reino de modo que la región perteneciente a cada príncipe compartiera una frontera de longitud no nula con las regiones pertenecientes a los otros cuatro príncipes? Möbius pasó el rompecabezas a sus estudiantes como un ejercicio. Pero en la lección siguiente él se disculpó por haberles pedido realizar lo imposible. Con esto quería decir que él podía *demostrar* que era imposible[31].

[31] Si se permite que las fronteras sean muy complicadas, no como en un mapa sino mucho más sinuosas, entonces tantos países como uno quiera pueden compartir una «frontera» común. Una construcción llamada los Lagos de Wada demuestra este resultado contraintuitivo. Véase:

Es difícil abordar este rompecabezas de forma geométrica, porque las formas de las regiones y cómo están dispuestas podría en principio ser muy complicada. El progreso depende de una gran simplificación: todo lo que realmente importa es qué regiones son adyacentes a cuales y cómo están dispuestas las fronteras comunes con relación a cada una de ellas. Esto es información topológica, independiente de las formas exactas. Puede representarse de una manera clara y simple conocida como un grafo, o, en nuestros días, una red, que es un término más evocador.

Una red es un concepto de una simplicidad abrumadora: un conjunto de vértices, representados por puntos, algunos de los cuales están conectados por aristas, dibujadas como líneas. Tomemos un mapa cualquiera, tal como el de la Figura 10 (*izquierda*). Para convertirlo en una red colocamos un punto dentro de cada región, Figura 10 (*centro*). Cuando quiera que dos regiones tienen un segmento de frontera común, dibujamos una línea entre los correspondientes puntos, que atraviesa dicho segmento. Si hay varios segmentos fronterizos comunes independientes, cada uno tiene su propia línea. Hagamos esto para todas las regiones y todos los segmentos fronterizos comunes, de tal modo que las líneas no se corten unas a otras, o a sí mismas, y solo se encuentren en los puntos. Luego prescindimos del mapa original y retenemos solo los

http://en.wikipedia.org/wiki/Lakes_of_Wada

puntos y las líneas. Estas forman la red dual del mapa, Figura 10 (*derecha*[32]).

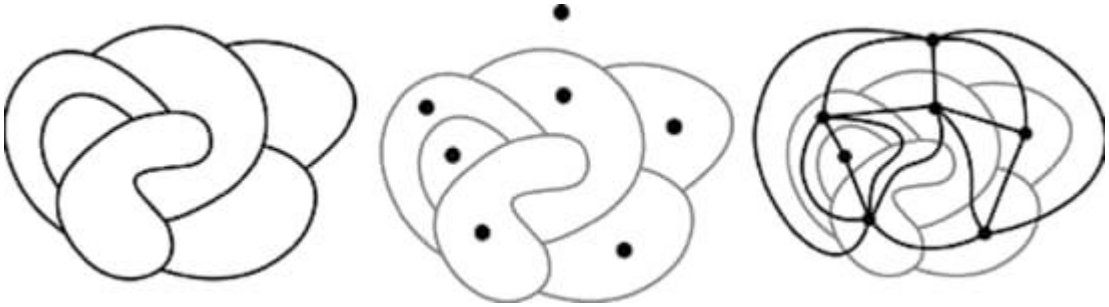


Figura 10. Izquierda: Un mapa. Centro: Se coloca un punto en cada región. Derecha: Se conectan puntos a través de las fronteras para formar la red dual (solo las líneas negras y los puntos).

Se utiliza la palabra «dual» porque el procedimiento toma regiones, líneas y puntos (uniones entre regiones del mapa) y los convierte en puntos, líneas y regiones. Una región en el mapa corresponde a un punto en la red dual. Un segmento fronterizo en el mapa corresponde a una línea en la red dual; no la misma línea, sino una que cruza la frontera y une los puntos correspondientes. Un punto en el mapa donde se juntan tres o más regiones corresponde a una región en la red dual acotada por un lazo cerrado de líneas. De modo que la propia red dual es un mapa, porque las líneas encierran regiones, y resulta que el dual del dual es el mapa original, salvo algunas cuestiones técnicas que excluyen puntos y líneas innecesarias.

[32] El término técnico es «grafo dual», porque tradicionalmente se utilizaba «grafo» en lugar de «red». Pero «red» se está haciendo habitual, es más evocador y evita la confusión con otros usos de la palabra «grafo».

El problema de los cinco príncipes puede reinterpretarse utilizando la red dual: ¿es posible unir cinco puntos del plano por líneas, sin cruces? La respuesta es «no», y la clave es la fórmula de Euler, que afirma que si un mapa en el plano consta de C caras (regiones), A aristas (líneas) y V vértices (puntos), entonces $C + V - A = 2$. Aquí contamos el resto del plano, fuera de la red, como una gran región. Esta fórmula era uno de los primeros indicios de que consideraciones topológicas podían ser dignas de investigar, y reaparecerá en el capítulo 10.

La demostración de que el rompecabezas de los príncipes indios es imposible empieza suponiendo que existe una solución, y de ello deduce una contradicción. Cualquier solución tendrá $V = 5$, el número de puntos. Puesto que cada par de puntos está unido por una línea, y hay 10 pares, $A = 10$. El teorema de Euler implica que $C = A - V + 2 = 7$. Las regiones de la red dual están rodeadas por lazos cerrados de líneas, y solo una línea une cualquier par de puntos; por consiguiente estos lazos deben contener al menos 3 líneas. Puesto que hay 7 regiones, eso hace al menos 21 líneas… excepto que toda línea está siendo contada dos veces porque separa dos regiones. De modo que hay al menos 10½ líneas. El número de líneas es un entero, de modo que de hecho debe haber al menos 11 líneas. Sin embargo, ya sabemos que hay 10 líneas. Esto es una contradicción lógica, y demuestra que no existe una red semejante. El rey no puede dividir su tierra de la forma prescrita.

El aspecto alentador de este argumento es que métodos topológicos elegantes nos permiten demostrar algo específico e interesante sobre

mapas. Sin embargo, contrariamente a un malentendido común, que De Morgan parece haber compartido, la imposibilidad de resolver el rompecabezas de los cinco príncipes indios no demuestra el teorema de los cuatro colores. Una demostración puede ser falsa incluso si su conclusión es correcta, o no se sabe que sea incorrecta. Si en algún lugar de una presunta demostración yo encuentro un triángulo con cuatro lados, puedo dejar de leer, porque la demostración es errónea. No importa lo que suceda después de eso o cuál sea la conclusión. Nuestra respuesta al rompecabezas de los príncipes indios muestra que una manera concreta de refutar el teorema de los cuatro colores no funciona. Sin embargo, eso no implica que no pueda funcionar ninguna otra manera de refutarlo. Potencialmente podría haber muchos obstáculos a 4-colorear un mapa (de ahora en adelante utilizaré este término en lugar del farragoso «colorear el mapa con cuatro colores»). La existencia de cinco regiones todas ellas adyacentes es simplemente uno de estos obstáculos. Por lo que sabemos, podría haber un mapa muy complicado con 703 regiones, tal que por más que se 4-coloreen 702 de ellas, la región final siempre necesita un quinto color. Esa región tendría que colindar al menos con otras cuatro, pero eso es perfectamente factible y no requiere una ordenación de príncipe indio. Si existiera un mapa así, demostraría que cuatro colores no son suficientes. Cualquier demostración tiene que descartar ese tipo de obstáculos. Y ese enunciado es válido incluso si yo no le muestro —no puedo mostrarle— un ejemplo explícito de tal obstáculo.

Durante un tiempo el teorema de los cuatro colores parecía haberse hundido sin dejar rastro, pero volvió a salir a la superficie en 1878 cuando Arthur Cayley lo mencionó en una reunión de la Sociedad Matemática de Londres. Pese a su nombre, esta organización representaba al conjunto de las matemáticas británicas (o al menos inglesas) y su fundador era De Morgan. Cayley preguntó si alguien había obtenido una solución. Su pregunta fue publicada poco después en la revista científica *Nature*. Un año más tarde él escribió un artículo más extenso para los *Proceedings of the Royal Geographical Society*[33]. Presumiblemente parecía un lugar lógico donde colocar el artículo, porque el problema trata ostensiblemente de mapas. Quizá incluso se le pidió que lo enviara. Pero en realidad no era una elección razonable, porque ningún cartógrafo iba a tener una razón para querer saber la respuesta, aparte de la curiosidad ociosa. Lamentablemente, la elección de revista significó que pocos matemáticos fueran conscientes de la existencia del artículo. Fue una lástima, porque Cayley explicaba por qué el problema podría ser difícil.

En el capítulo 1 dije que una demostración se parece un poco a una batalla. Los militares reconocen una diferencia entre táctica y estrategia. La táctica es como se ganan las escaramuzas locales; la estrategia establece la estructura general de la campaña. La táctica incluye movimientos detallados de tropas; la estrategia incluye planes generales, con lugar para muchas decisiones tácticas

[33] Hasta hace poco se pensaba que la nota en *Nature* era la última referencia en prensa al problema durante casi un siglo, pero el historiador de las matemáticas Robin Wilson localizó este artículo posterior de Cayley.

diferentes en cualquier etapa. El artículo de Cayley era escaso en táctica, pero un vago indicio de una estrategia que, con el tiempo, abrió de par en par el problema de los cuatro colores. Él observaba que añadir regiones de una en una no funcionaba si se seguía la línea obvia de razonamiento. Pero quizá funcionaría si se encuentra una línea de razonamiento menos obvia.

Supongamos que se toma un mapa y se elimina una región —digamos que fusionándola con una vecina, o contrayéndola hasta un punto—. Supongamos que el mapa resultante puede ser 4-coloreado. Ahora recuperamos la región original. Si tenemos suerte, sus vecinos podrían utilizar solo tres colores. Entonces todo lo que hay que hacer es colorearla utilizando el cuarto. El punto de Cayley era que este procedimiento podría no funcionar, porque los vecinos de la región final podrían utilizar cuatro colores distintos. Pero eso no significa que estemos atascados. Hay dos maneras de burlar este obstáculo: quizá hemos escogido la región equivocada, o quizá hemos escogido la manera equivocada de colorear el mapa más pequeño.

Siguiendo con hipótesis no demostradas (esta es una manera muy efectiva de obtener ideas en investigación, aunque en algún momento hay que confirmarlas), supongamos que siempre se puede hacer un arreglo parecido. Eso nos dice que un mapa siempre puede ser 4-coloreado con tal de que algún mapa más pequeño pueda ser 4-coloreado. Quizá esto no parezca un avance: ¿cómo sabemos que el mapa más pequeño puede ser 4-coloreado? La respuesta es que el mismo procedimiento se aplica al mapa más

pequeño, lo que lleva a un mapa aún más pequeño... y así sucesivamente. Por último se llega a un mapa tan pequeño que solo tiene cuatro regiones, y entonces se sabe que puede ser 4-coloreado. Invirtamos ahora los pasos, coloreando mapas ligeramente más grandes en cada etapa... y finalmente regresamos al mapa original.

Esta línea de razonamiento se denomina «demostración por inducción matemática». Es un método estándar con una formulación más técnica, y la lógica que hay tras ello puede hacerse rigurosa. La estrategia de demostración propuesta por Cayley se hace más transparente si el método se reformula utilizando un concepto lógicamente equivalente: el de un criminal mínimo. En este contexto, un criminal es cualquier mapa hipotético que no puede ser 4-coloreado. Un mapa tal es mínimo si cualquier mapa con un número menor de países *puede* ser 4-coloreado. Si existe un criminal, debe ser un criminal mínimo: simplemente escogemos un criminal con el mínimo número de regiones posible. Por consiguiente, si no existe un criminal mínimo, entonces no existen criminales. Y si no hay criminales, el teorema de los cuatro colores debe ser cierto.

El procedimiento de inducción se reduce a esto. Supongamos que podemos demostrar que 4-colorear un criminal mínimo es siempre posible, con tal de que pueda 4-colorearse un mapa relacionado más pequeño. Entonces el criminal mínimo no puede ser realmente un criminal. Puesto que el mapa es mínimo, *todos* los mapas más pequeños pueden ser 4-coloreados, y dado lo que hemos supuesto que puede demostrarse, lo mismo es cierto para el mapa original.

Por consiguiente no hay criminales mínimos, luego no hay criminales. Esta idea desplaza el foco del problema desde todos los mapas a solo los hipotéticos criminales mínimos y a especificar un procedimiento de reducción, esto es, una manera sistemática de convertir un 4-coloreado de un mapa relacionado más pequeño en un 4-coloreado del mapa original.

¿Por qué molestarnos con criminales mínimos y no con simples criminales? Es una cuestión de técnica. Incluso si inicialmente no sabemos que existen criminales, una de las características paradójicas pero útiles de esta estrategia es que podemos decir muchas cosas sobre qué aspecto tendrían los criminales mínimos si existieran.

Esto requiere la capacidad de pensar lógicamente sobre proposiciones hipotéticas, una habilidad vital para cualquier matemático. Para dar una idea del proceso, demostraré el teorema de los *seis* colores. Para hacerlo, tomamos prestado un truco del rompecabezas de los cinco príncipes y reformulamos todo en términos de la red dual, en la que las regiones se convierten en puntos. El problema de los cuatro colores equivale entonces a una pregunta diferente: dada una red en el plano cuyas líneas no se cruzan, ¿es posible 4-colorear los *puntos* de modo que dos puntos unidos por una línea tengan siempre colores diferentes? La misma reformulación se aplica a cualquier número de colores.

Para ilustrar la potencia de los criminales mínimos, voy a utilizarlos para demostrar que cualquier red plana puede ser 6-coloreada. Una vez más, la principal herramienta técnica es la fórmula de Euler.

Dado un punto en la red dual, definimos sus vecinos como aquellos puntos que están unidos a él por una línea. Un punto puede tener muchos vecinos o solo unos pocos. Puede demostrarse que la fórmula de Euler implica que algunos puntos deben tener pocos vecinos. Más exactamente, en una red plana es imposible que todos los puntos tengan seis o más vecinos. He introducido una demostración de esto en las notas para no interrumpir el flujo de ideas[34]. Este hecho proporciona la palanca necesaria para empezar a desmontar el problema. Consideremos un hipotético criminal mínimo para el teorema de seis colores. Este es una red que no puede ser 6-coloreada pero toda red más pequeña sí puede ser 6-coloreada. Ahora demuestro que este mapa no puede existir. Por la consecuencia anterior de la fórmula de Euler, contiene al menos un

---

[34] Trabajando en la red dual, sea C el número de caras (incluida una gran cara que rodea la red entera), A el número de aristas y V el número de vértices. Podemos suponer que cada cara en la red dual tiene al menos tres aristas; si tiene una cara con solo dos aristas entonces corresponde a un vértice «superfluo» de la red original en donde se encuentran solo dos aristas. Este vértice puede eliminarse y las dos aristas se unen. Cada arista limita dos caras, y cada cara tiene al menos tres aristas, de modo que $A \geq 3C/2$, o lo que es equivalente $2A/3 \geq C$. Por el teorema de Euler $C + V - A = 2$, de modo que $2A/3 + V - A \geq 2$, lo que implica que

$$12 + 2A \leq 6V$$

Supongamos que Am es el número de vértices con m vecinos. Entonces $A2$, $A3$, $A4$ y $A5$ son cero. Por lo tanto

$$A = A6 + A7 + A8 + \ldots$$

Puesto que cara arista une dos vértices,

$$2A = 6V6 + 7V7 + 8V8 + \ldots$$

Sustituyendo esto en la desigualdad obtenemos

$$12 + 6V6 + 7V7 + 8V8 + \ldots \leq 6V6 + 6V7 + 6V8 + \ldots$$

de modo que

$$12 + V7 + 2V8 + \ldots \leq 0$$

que es imposible.

punto con cinco o menos vecinos. Borramos temporalmente este punto y las líneas que lo unen a sus vecinos. La red resultante tiene menos puntos, de modo que por minimalidad puede ser 6-coloreada. (Aquí es donde nos quedamos atascados a menos que nuestro hipotético criminal sea mínimo). Ahora recuperamos el punto y las líneas eliminadas. Ese punto tiene a lo sumo cinco vecinos, de modo que siempre hay un sexto color. Lo utilizamos para colorear el punto que habíamos eliminado. Ahora hemos 6-coloreado con éxito nuestro criminal mínimo... pero eso contradice su criminalidad. Así que no existen criminales mínimos en el teorema de los seis colores, y eso implica que el teorema de los seis colores es cierto.

Esto es alentador. Hasta ahora, por lo que sabemos, algunos mapas podrían necesitar 20 colores, o 703, o millones. Ahora sabemos que mapas como ese no son más reales que el puchero lleno de monedas de oro en el extremo del arco iris. Un número concreto y limitado de colores funciona decididamente para *cualquier* mapa. Este es un triunfo genuino para criminales mínimos, y animó a los matemáticos a reforzar el argumento con la esperanza de reemplazar seis colores por cinco o, si uno fuera realmente ingenioso, cuatro.

Todos los criminales necesitan abogados. Un abogado de los tribunales llamado Alfred Kempe estaba en la reunión en la que Cayley mencionó el problema de los cuatro colores. Él había estudiado matemáticas con Cayley en Cambridge y su interés por ellas seguía intacto. En menos de un año Kempe se convenció de

que había resuelto el problema y publicó su solución en 1879 en el recientemente fundado *American Journal of Mathematics*. Un año más tarde publicó una demostración simplificada, que corregía algunos errores de la primera. Él señalaba que

> *una alteración muy pequeña en una parte de un mapa puede hacer necesario recolorearlo por completo. Tras una ardua investigación, he conseguido... dar con el punto débil, que resultó fácil de atacar.*

Reinterpretaré las ideas de Kempe en términos de la red dual. Una vez más, él partió de la fórmula de Euler y la consecuente existencia de un punto con tres, cuatro o cinco vecinos. (Un punto con dos vecinos se encuentra en la mitad de una línea, y no aporta nada a la red o el mapa: puede omitirse sin problemas).

Si hay un punto con tres vecinos, el procedimiento que utilicé para demostrar el teorema de los seis colores se aplica cuando solo hay cuatro colores. Eliminamos el punto y las líneas que se encuentran en él, 4-coloreamos el resultado, recuperamos el punto y las líneas, utilizamos un color de reserva para el punto. Por consiguiente podemos suponer que ningún punto tiene tres vecinos.

Si hay un punto con cuatro vecinos la táctica anterior falla, porque podría no haber disponible un color de reserva. Kempe ideó una manera ingeniosa de tratar este obstáculo. Borrar ese punto en cualquier caso pero, después de hacerlo, cambiar el coloreado del mapa más pequeño resultante de modo que dos de los cuatro vecinos tengan el mismo color. Tras este cambio, los vecinos del punto borrado utilizan a lo sumo tres colores, lo que deja uno de

reserva para el punto eliminado. La idea básica del esquema de recoloreado de Kempe es que dos de los puntos vecinos deben tener colores diferentes —digamos rojo y azul, siendo los otros colores verde y amarillo—. Si ambos son verdes o ambos son amarillos, el otro color queda disponible para el punto eliminado. Así que podemos suponer que uno es verde y el otro amarillo. Ahora encontramos todos los puntos que pueden ser conectados con el azul por una secuencia de líneas, utilizando solamente puntos azules y rojos. Llamamos a esto una cadena de Kempe azul-roja[35]. Llamamos a esto una cadena de Kempe azul-r. Por definición, cualquier punto que no está en la cadena de Kempe, pero es vecino de un punto en la cadena, es o verde o amarillo, porque un vecino azul o rojo ya estaría en la cadena. Habiendo encontrado dicha cadena, observamos que intercambiar los dos colores azul y rojo para todos los puntos dentro de la cadena produce otro coloreado de la red, que sigue satisfaciendo la condición clave de que puntos adyacentes tienen colores diferentes (véase Figura 11).

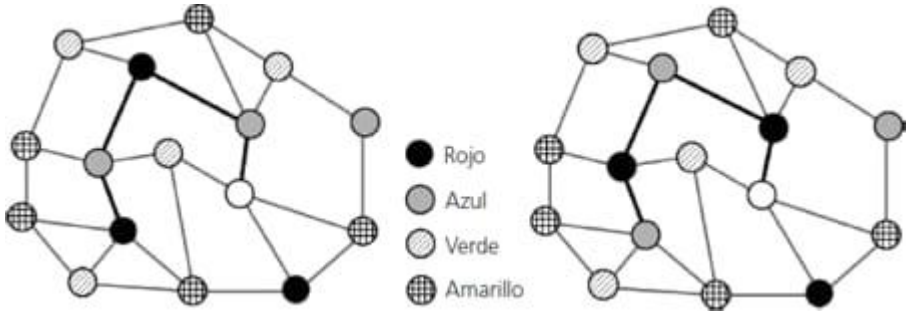


Figura 11. Intercambio de colores en una cadena de Kempe (líneas negras gruesas) con un punto de grado 4 (blanco) que tiene vecinos

[35] «Cadena» es equívoco, puesto que sugiere una secuencia lineal. Una cadena de Kempe puede contener lazos y puede ramificarse.

de los cuatro colores. Izquierda: Colores originales. Derecha: Con los colores intercambiados, el azul está disponible para el punto blanco.

Si el vecino rojo de nuestro punto original no está en esta cadena azul-roja, hagamos dicho cambio. El vecino azul de los puntos originales se vuelve rojo; el vecino rojo sigue siendo rojo. Ahora los vecinos del punto original utilizan a lo sumo tres colores diferentes: rojo, verde y amarillo. Esto deja el azul para el punto original, y ya está hecho. Sin embargo, la cadena azul-roja podría enrollarse y unirse con el vecino azul. Si es así, dejamos sola la cadena azul y roja, y en su lugar utilizamos el mismo truco para los vecinos amarillo y verde del punto original. Empezamos con el verde y formamos una cadena de Kempe verde-amarilla. Esta cadena *no puede* unirse con el vecino amarillo, porque la cadena azul-roja anterior se interpone en el camino. Intercambiamos amarillo y verde, y está hecho.

Eso deja un último caso, cuando no hay puntos con tres o cuatro vecinos pero uno al menos tiene cinco vecinos. Kempe propuso una regla de recoloreado similar aunque más complicada, que parecía resolver también ese caso. Conclusión: el teorema de los cuatro colores es verdadero, y Kempe lo había demostrado. Incluso llegó a los medios de comunicación: *The Nation*, una revista norteamericana, mencionó la solución en su sección de recensiones. Parecía que la demostración de Kempe había enterrado el problema. Para la mayoría de los matemáticos era cosa hecha. Peter Guthrie Tait continuó publicando artículos sobre el problema, buscando una

demostración más sencilla; esto le llevó a algunos descubrimientos útiles, pero la demostración más sencilla le eludía.

Entra en acción Percy Heawood, un profesor de matemáticas en la Universidad de Durham conocido por el apodo «Pussy» gracias a su magnífico bigote. Cuando era estudiante en Oxford había sabido del problema de los cuatro colores por Henry Smith, el catedrático de Geometría. Smith le dijo que el teorema, aunque posiblemente cierto, no estaba demostrado, así que Heawood tenía una meta. En el camino dio con el artículo de Kempe, y trató de entenderlo. Publicó el resultado en 1889 como el «Teorema de colorear mapas», lamentando que el propósito de su artículo era «más destructivo que constructivo, pues se mostrará que hay un defecto en la demostración ahora aparentemente admitida». Kempe había cometido un error.

Era un error sutil: ocurría en el método de recoloreado cuando el punto que se borraba tenía cinco vecinos. El esquema de Kempe podía en ocasiones cambiar el color de algún punto como repercusión de cambios posteriores. Pero Kempe había supuesto que una vez que se había cambiado el color de un punto, no volvía a cambiar. Heawood encontró una red para la que el esquema de recoloreado de Kempe no funcionaba, de modo que su demostración era defectuosa. Kempe reconoció rápidamente el error y añadió que él «no había conseguido remediar el defecto». El teorema de los cuatro colores estaba de nuevo abierto.

Heawood sacó de la debacle algo de consuelo para Kempe: su método demostraba satisfactoriamente el teorema de los cinco

colores. Heawood también desarrolló dos generalizaciones del problema: los imperios, en los que las regiones pueden constar de varias piezas inconexas, todas las cuales requieren el mismo color; y mapas en superficies más complicadas. La pregunta análoga en una esfera tiene la misma respuesta que para el plano. Imaginemos un mapa en una esfera y lo rotamos hasta que el polo norte está en alguna parte dentro de una región. Si se borra el polo norte se puede abrir la esfera punteada para obtener un espacio que es topológicamente equivalente al plano infinito. La región que contiene al polo se convierte en la región infinitamente grande que rodea al resto del mapa. Pero hay otras superficies más interesantes. Entre ellas está el toro, que tiene la forma de un donut con un agujero (véase Figura 12, *izquierda*).

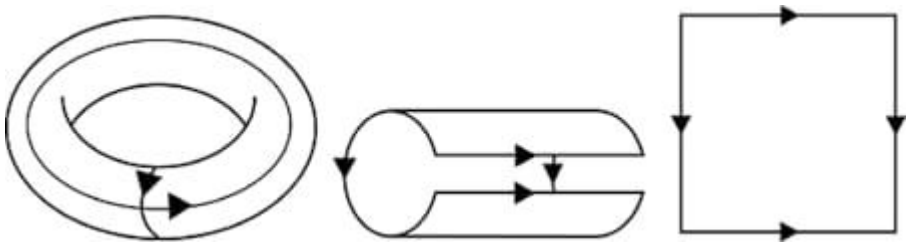


Figura 12. Cortando un toro abierto y desplegándolo para formar un cuadrado.

Hay una forma útil de visualizar el toro, que a menudo hace la vida más sencilla. Si cortamos el toro a lo largo de dos curvas cerradas (Figura 12, *centro*), podemos desplegarlo en un cuadrado (Figura 12, *derecha*). Esta transformación cambia la topología del toro, pero podemos sortearlo «identificando» lados opuestos del cuadrado. En

efecto (y una definición rigurosa hace esta idea precisa), acordamos tratar puntos correspondientes en estos lados como si fueran idénticos. Para ver cómo va esto, invirtamos la secuencia de imágenes. El cuadrado se enrolla, y los lados opuestos realmente se empalman. Ahora llega la parte más ingeniosa. En realidad no se necesita enrollar el cuadrado y unir los lados correspondientes. Simplemente podemos trabajar con el cuadrado plano, siempre que tengamos en mente la regla para identificar los lados. Todo lo que hagamos con el toro, como dibujar curvas en el mismo, tiene una correspondiente construcción precisa en el cuadrado.

Heawood demostró que siete colores son necesarios y suficientes para colorear cualquier mapa en un toro. La Figura 13 (*izquierda*) muestra que siete son necesarios, utilizando un cuadrado para representar el toro como se acaba de describir. Observemos cómo se unen las regiones en lados opuestos. Hay superficies como un toro, pero con más agujeros (véase Figura 13, *derecha*). El número de agujeros se denomina el género y se denota por la letra *g*. Heawood conjeturó una fórmula para el número de colores requerido en un toro con *g* agujeros cuando $g \geq 1$: es el mínimo número entero menor o igual que

$$\frac{7+\sqrt{48g+1}}{2}$$

Cuando *g* va de 1 a 10, está fórmula da los números:

7 8 9 10 11 12 12 13 13 14

El número de colores especificado por la fórmula crece más lentamente que el género, y a menudo no supone ninguna diferencia si se añade un agujero extra al toro. Esto es una sorpresa, porque todo agujero extra proporciona más libertad para inventar mapas complicados.

Heawood no sacó esta fórmula del aire. Surgió al generalizar la manera en que he demostrado el teorema de seis colores en el plano. Él pudo demostrar que este número de colores siempre es suficiente. La gran pregunta, durante muchos años, era si este número puede hacerse más pequeño. Ejemplos para valores pequeños del género sugerían que la estimación de Heawood es la mejor posible. En 1968, después de una larga investigación, Gerhard Ringel y John W. T. (Ted) Youngs completaron los detalles finales en una demostración de que esto es correcto, basados en su propio trabajo y en el de otros. Sus métodos son combinatorios, basados en tipos de redes especiales, y suficientemente complicados para llenar todo un libro [36].

---

[36] La demostración completa se da en Gerhard Ringel, *Map Color Theorem*, Springer, 1974. Se divide en doce casos, dependiendo de si el género es de la forma $12k$, $12k + 1$..., $12k + 11$. Llamemos a estos casos 0-11. Con un número finito de excepciones, los casos se resolvieron como sigue:

Caso 5: Ringel, 1954
Casos 3, 7 y 10: Ringel en 1961.
Casos 0 y 4: C. M. Ferry, Lloyd Welch y Youngs en 1963.
Caso 1: W. Gustin y Youngs en 1964.
Caso 9: Gustin en 1965.
Caso 6: Youngs en 1966.
Casos 2, 8 y 11: Ringel y Youngs en 1967.

Las excepciones eran los géneros 18, 20, 23 (resuelto por Yves Mayer en 1967) y 30, 35, 47, 659 (resuelto por Ringel y Youngs en 1968). Ellos también trataron el problema análogo para

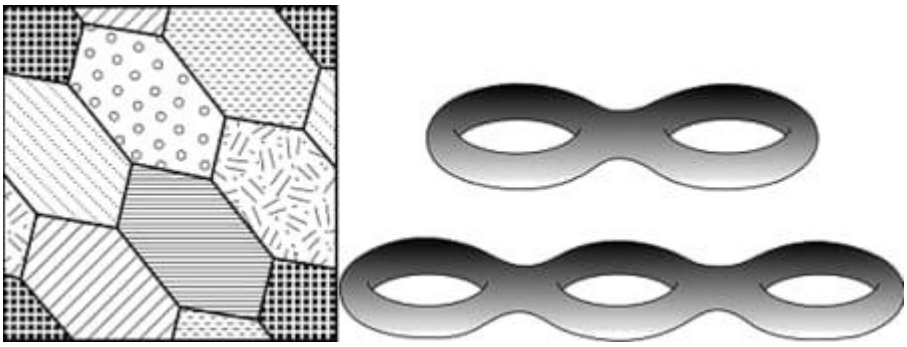


Figura 13. Izquierda: Mapa de siete colores en un toro. El toro está representado como un cuadrado cuyos lados opuestos están «enrollados» idealmente de modo que se juntan. Se requiere que las regiones del mapa encajen a través de los bordes correspondientes. Derecha: Toros con dos y tres agujero.

Cuando $g = 0$, es decir, para mapas en una esfera, la fórmula de Heawood da cuatro colores, pero su demostración de suficiencia no funciona en una esfera. Pese a avances impresionantes para superficies con al menos un agujero, el problema de cuatro colores original estaba aún abierto para cualquiera. Los pocos matemáticos que estaban dispuestos a dedicar serios esfuerzos a la cuestión se prepararon para lo que, en términos bélicos, iba a ser sin duda un largo asedio. El problema era un castillo fuertemente defendido. Ellos esperaban construir máquinas de asedio cada vez más poderosas y seguir abriendo brechas hasta que cayeran los muros del castillo. Y así lo hicieron, pero el castillo no cayó. Sin embargo, los atacantes acumularon lentamente una gran cantidad de

superficies de una cara (como la banda de Möbius pero sin bordes), que también había abordado Heawood.

información sobre cómo no resolver el problema y sobre los tipos de obstáculos que parecían inevitables. A partir de estos fallos empezó a surgir una ambiciosa estrategia. Era una ampliación natural de los métodos de Kempe y de Heawood, y llegó en tres partes. Las enunciaré utilizando la red dual, el punto de vista estándar en nuestros días:

1. Considerar un criminal mínimo.
2. Encontrar una lista de configuraciones inevitables: redes más pequeñas con la propiedad de que cualquier criminal mínimo debe contener algo que está en la lista.
3. Demostrar que cada una de las configuraciones inevitables es reducible. Es decir: si una red más pequeña, obtenida eliminando la configuración inevitable, puede ser 4-coloreada, entonces estos colores pueden redistribuirse de modo que cuando se recupera la configuración inevitable, el 4-coloreado de la red más pequeña se extiende a la red entera.

Juntando estos tres pasos podemos demostrar que no existe un criminal mínimo. Si existiera, contendría una configuración inevitable. Pero el resto de la red es más pequeño, de modo que la minimalidad implica que puede ser 4-coloreado. La reducibilidad implica ahora que la red original puede ser 4-coloreada. Esto es una contradicción.

En estos términos, Kempe había encontrado correctamente una lista de configuraciones inevitables: un punto en el que se juntan tres líneas, uno en el que se juntan cuatro y uno en el que se juntan cinco (veáse Figura 14). Él también había demostrado

correctamente que las dos primeras son reducibles. Su error estaba en su demostración de que la tercera configuración es reducible. No lo es. Propuesta: reemplazar esta mala configuración por una lista más larga, asegurando que la lista sigue siendo inevitable. Hacerlo de tal manera que cada configuración en la nueva lista es reducible. Es decir: buscar una lista inevitable de configuraciones reducibles. Si se consigue, se ha demostrado el teorema de los cuatro colores.

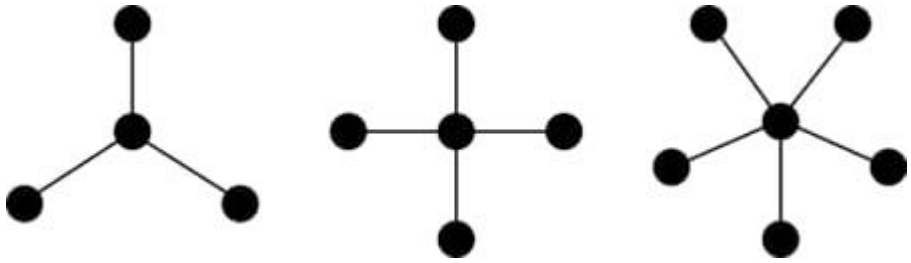


Figura 14. Lista de Kempe de configuraciones inevitables.

Podría no haber tal lista pero esta estrategia bien vale un intento, y nadie tenía ideas mejores. Tiene una delicada tensión interna, no obstante. Por una parte, cuanto más larga es la lista, más probabilidades tiene de ser inevitable, lo que es bueno. Por otra parte, cuanto más larga es la lista, menos probable es que toda configuración en ella sea reducible. Basta con que una sola no lo sea para que la demostración entera colapse, y este peligro se hace más agudo a medida que la lista crece, lo que es malo. Por una tercera parte… una lista más larga proporciona más oportunidades para escoger configuraciones reducibles, lo que es bueno. Por una cuarta parte, aumenta el trabajo necesario para demostrar la reducibilidad, lo que es malo. Y por una quinta parte, en cualquier caso no había nuevos métodos para hacerlo, lo que era peor.

Son cosas de este tipo las que hacen grande un gran problema.
Así que durante un tiempo la parte singular del castillo fue cediendo ocasionalmente, pero su pérdida no suponía la más mínima diferencia para la solidez de la fortaleza. Mientras tanto las matemáticas de la corriente principal bostezaban, si es que prestaban alguna atención. Pero alguien estaba construyendo un ariete mejor, y su nombre era Heinrich Heesch. Su gran aportación fue una manera sistemática de demostrar que una configuración es reducible. Él la llamaba «descargar», y era algo parecido a imaginar que los puntos en la red llevan carga eléctrica y permitir que la electricidad fluya de un punto a otro.
Incluso con este método, encontrar a mano un conjunto inevitable de configuraciones reducibles sería una tarea abrumadora. Las configuraciones individuales serían probablemente bastante pequeñas, pero tendría que haber muchas de ellas. Heesch insistió, y en 1948 dio una serie de conferencias sugiriendo que serían necesarias unas diez mil configuraciones. Para entonces ya había demostrado que quinientos candidatos eran reducibles. Entre la audiencia había un joven llamado Wolfgang Haken, quien más tarde dijo que en realidad no había entendido mucho de las lecciones de Heesch en esa época, aunque se le habían quedado en la cabeza algunos de los puntos principales. Haken pasó a estudiar topología e hizo un avance importante en la teoría de nudos. Esto le animó a trabajar en la conjetura de Poincaré (véase capítulo 10). Para una línea de ataque particular, clasificó las posibilidades en doscientos

casos, resolvió 198 de ellos y especuló con los dos restantes durante trece años. En ese momento lo dejó, y en su lugar empezó a trabajar en el problema de los cuatro colores. Es evidente que a Haken le gustaban los problemas difíciles, pero le preocupaba que algo similar pudiera haber sucedido con las diez mil configuraciones de Heesch. Imaginemos que tratamos acertadamente 9998 y nos quedamos atascados en las dos últimas. Por ello, en 1967 Haken invitó a Heesch a visitar la Universidad de Illinois, donde él trabajaba, para pedirle consejo.

En aquellos días los ordenadores empezaban a hacerse útiles para las matemáticas reales, pero eran máquinas enormes situadas en algún edificio central, no objetos que descansaban en la mesa o dentro de un maletín. Haken se preguntaba si podrían ser de ayuda. Heesch ya había tenido la misma idea e hizo una estimación aproximada de la complejidad del problema. Esta indicaba que el mejor ordenador del que podía disponer no estaba a la altura de la tarea. Illinois tenía un ordenador mucho más potente, un ILLIAC-IV, de modo que Haken solicitó tiempo. Pero el superordenador no estaba listo, y por ello se le dijo que lo intentase en el Cray 600 del Laboratorio de Brookhaven en Long Island. El director del centro de computación del laboratorio era Yoshio Shimamoto, que llevaba mucho tiempo fascinado con el problema de los cuatro colores; esto fue un golpe de suerte que dio a Heesch y Haken acceso a la máquina.

El ordenador respondió a las expectativas, pero Haken empezó a preguntarse si podía utilizarse de forma más eficiente. Ellos estaban

generando montones de configuraciones reducibles y esperaban reunir una lista inevitable, pero esa estrategia gastaba mucho tiempo en configuraciones potenciales que resultaron no ser reducibles. ¿Por qué no hacerlo al revés: hacer de la inevitabilidad el objetivo principal y comprobar más tarde la reducibilidad? Por supuesto, habría que utilizar configuraciones que tuvieran una alta probabilidad de ser reducibles, pero parecía una mejor manera de avanzar. Para entonces, sin embargo, el Cray de Brookhaven estaba siendo utilizado en cosas más importantes. Peor aún, varios expertos dijeron a Haken que los métodos que él quería utilizar no podían traducirse en programas de ordenador. Él los creyó, y dio una conferencia diciendo que el problema no podía resolverse sin ordenadores pero ahora parecía que tampoco podía resolverse con ordenadores. Había decidido abandonar.

Entre la audiencia había un experto programador, Kenneth Appel, quien dijo a Haken que probablemente los presuntos expertos solo estaban tratando de disuadirle porque el programa requeriría mucho trabajo y el resultado era muy incierto. En opinión de Appel no había ningún problema matemático que no pudiera ser programado. La cuestión crucial era si el programa llegaría a alguna parte en un tiempo razonable. Ambos unieron sus fuerzas. La estrategia evolucionó a medida que mejoras en el método de descarga provocaban cambios en el programa y las mejoras en el programa provocaban cambios en el método de descarga. Esto les llevó a un nuevo concepto: configuraciones «geográficamente buenas», que no contenían ciertas configuraciones molestas que

impedían la reducibilidad. La probabilidad de que una configuración semejante fuera reducible estaba muy mejorada, y la propiedad definitoria era fácil de comprobar. Appel y Haken decidieron demostrar con la teoría, en lugar de mediante ordenador, que había una lista inevitable de configuraciones geográficamente buenas. En 1974 lo habían conseguido.

Esto era alentador, pero ellos sabían lo que sin duda iba a suceder. Algunas de sus configuraciones geográficamente buenas resultarían no ser reducibles, de modo que tendrían que eliminarlas y reemplazarlas por una lista más larga y más complicada. El cálculo se estaba persiguiendo la cola y solo tendría éxito si la atrapaba. En lugar de perder años en una búsqueda infructuosa, hicieron algunas crudas estimaciones de cuánto tiempo podría llevar el proceso. Los resultados fueron moderadamente alentadores, de modo que el trabajo continuó. Teoría y computación se alimentaban y cambiaban una a otra. A veces, el ordenador parecía tener una mente propia, «descubriendo» propiedades útiles de las configuraciones. En ese momento la administración de la universidad compró para su propio uso un nuevo ordenador muy potente; más potente, de hecho, que los que estaban disponibles para los científicos de la universidad. Después de algunas protestas y preguntas mordaces, la mitad del tiempo de la máquina se puso a disposición de los científicos. La lista siempre cambiante de configuraciones inevitables de Appel y Haken se estabilizó en unas dos mil de ellas. En junio de 1976 el ordenador realizó su última comprobación de reducibilidad, y la demostración estaba completa.

La historia llegó a los medios de comunicación, empezando por *The Times*, y rápidamente se difundió por todo el mundo.

Aún tenían que asegurar que no había errores estúpidos, y para entonces otros equipos les pisaban los talones. En julio, Appel y Haken confiaban en que su método funcionaba y anunciaron oficialmente su demostración a la comunidad matemática. Para ello hicieron circular una prepublicación —una versión preliminar y de fácil distribución de un artículo del que se pretende una publicación posterior—. En esa época, lo normal era que pasaran entre uno y dos años antes de tener un artículo matemático en prensa. Para evitar un atasco en el progreso la profesión tenía que encontrar una forma más rápida para comunicar resultados importantes a la comunidad, y las prepublicaciones eran la forma de hacerlo. En nuestros días, las prepublicaciones se colocan en la web. Las prepublicaciones son siempre provisionales; la publicación definitiva requiere revisión por pares. Las prepublicaciones ayudan en este proceso porque cualquiera puede leerlos, buscar errores o mejoras, y decírselo a los autores. De hecho, la versión publicada suele diferir de modo notable de la prepublicación, precisamente por esta razón.

La demostración final necesitó mil horas de tiempo del ordenador e incluía 487 reglas de descarga; los resultados se publicaron en forma de dos artículos con un suplemento de 450 páginas que mostraba 1482 configuraciones. En esa época fue un *tour de force*.

Sin embargo, la reacción principal de la comunidad matemática más amplia fue de vago malestar. No por el resultado; no por el notable logro computacional. Lo que producía malestar era el

método. En los años setenta las demostraciones matemáticas eran cosas que se hacían a mano y se comprobaban a mano. Como dije en el capítulo 1, una demostración es una historia cuyo argumento convence de que el enunciado es verdadero. Pero esta historia no tenía un argumento. O si lo tenía, había un gran agujero en el centro:

> *Había una vez una bella conjetura. Su madre le dijo que nunca debía entrar en el bosque oscuro y peligroso. Pero un día la Pequeña Conjetura de los Cuatro Colores se escapó y se adentró en el bosque inevitable. Ella sabía que si cada configuración en el bosque fuera reducible, tendría una demostración, se convertiría en el Pequeño Teorema de los Cuatro Colores y sería publicada en una revista dirigida por el Príncipe Tom. Ella llegó a un ordenador cubierto de caramelo, en lo profundo del bosque, y dentro había un Lobo disfrazado de programador. Y el Lobo dijo: «Sí, todas son reducibles», y todos fueron felices.*

No, no funciona. Estoy frivolizando, pero el agujero en este cuento de hadas es el mismo agujero que hay en la demostración de Appel-Haken, o, al menos, lo que muchos matemáticos consideraban que era el agujero en la demostración. ¿*Cómo sabemos que el Lobo está en lo cierto*?

Ejecutamos nuestro propio programa de ordenador y vemos si coincide. Pero por muchas veces que lo hagamos, no tiene la misma aureola de autenticidad que, por ejemplo, mi demostración de que

no se puede cubrir con fichas de dominó un tablero de ajedrez recortado. No se puede captar en conjunto. Nadie podría comprobar a mano todos los cálculos aunque viviese mil millones de años. Peor aún, si se pudiera, nadie creería la respuesta. Los seres humanos cometen errores. En mil millones de años, cometen un montón de errores.

Los ordenadores, en general, no los cometen. Si un ordenador y un ser humano hacen un cálculo aritmético realmente complicado y discrepan, lo inteligente es apostar a favor del ordenador. Pero eso no es seguro. Un ordenador que está funcionado exactamente como se ha planeado puede cometer un error; por ejemplo, un rayo cósmico puede atravesar su memoria y cambiar un 0 por un 1. Eso puede precaverse haciendo de nuevo el cálculo; pero, lo que es más serio, los diseñadores pueden cometer errores. El chip Intel P5 Pentium tenía un error en sus rutinas de aritmética de punto flotante: si se le pedía que dividiera 4.195.835 por 3.145.727, respondía con 1,33373, cuando la respuesta correcta es 1,33382. Al parecer, se habían omitido cuatro entradas en una tabla[37]. Otras cosas que pueden fallar incluyen el sistema operativo del ordenador y errores en el programa del usuario.

Se ha dedicado mucha palabrería filosófica a la proposición de que la demostración asistida por ordenador de Appel-Haken cambió la naturaleza de la «demostración». Yo puedo ver lo que los filósofos pretenden, pero el concepto de demostración que utilizan quienes trabajan en matemáticas no es el que enseñamos a los estudiantes

[37] La notable historia de cómo se descubrió el error, y qué sucedió entonces, puede encontrarse en: http://en.wikipedia.org/wiki/Pentium_FDIV_bug

universitarios en las clases de lógica matemática. E incluso cuando se aplica ese concepto más formal, nada exige que la lógica de cada paso sea comprobada por un ser humano. Durante siglos los matemáticos han utilizado máquinas para la aritmética rutinaria. E incluso si un ser humano revisa una demostración línea por línea, sin encontrar errores, ¿cómo sabemos que no ha pasado por alto uno? La lógica perfecta e incuestionable es un ideal al que aspiramos. Los seres humanos imperfectos hacen lo mejor que pueden, pero nunca pueden eliminar todo elemento de incertidumbre.

En *Four Colours Suffice*, Robin Wilson pone el dedo en un aspecto sociológico clave de la reacción de la comunidad:

> *La audiencia se dividía en dos grupos: a los de más de cuarenta años no se les podía convencer de que una demostración por ordenador era correcta, mientras que a los de menos de cuarenta años no se les podía convencer de que una demostración que contiene setecientas páginas de cálculos a mano podía ser correcta.*

Si nuestras máquinas nos superan en algunas cosas, tiene sentido utilizar máquinas. Las *técnicas* de demostración pueden cambiar, pero lo hacen continuamente; eso se llama «investigación». El concepto de demostración no se altera radicalmente si algunos pasos los hace un ordenador. Una demostración es una historia; una demostración asistida por ordenador es una historia que es demasiado larga para ser contada con todo detalle, de modo que hay

que hacer un resumen ejecutivo y un enorme apéndice automatizado.

Desde el trabajo pionero de Appel y Haken, los matemáticos se han acostumbrado a la ayuda del ordenador. Siguen *prefiriendo* las demostraciones que se basan solamente en la potencia del cerebro humano, pero la mayoría de ellos no hacen de ello un requisito. En los años noventa, sin embargo, había todavía un cierto malestar justificable ante la demostración de Appel-Haken. Por ello, en lugar de volver a comprobar el trabajo, algunos matemáticos decidieron rehacer toda la demostración, sacando ventaja de los nuevos avances teóricos y los ordenadores muy mejorados. En 1994 Neil Robertson, Daniel Sanders, Paul Seymour y Robin Thomas prescindieron de todo lo que había en el artículo de Appel y Haken excepto la estrategia básica. En menos de un año habían encontrado un conjunto inevitable de 633 configuraciones, de cada una de las cuales podía demostrase que era reducible utilizando solamente 32 reglas de descarga. Esto era mucho más sencillo que las 1482 configuraciones y 487 reglas de descarga. Los ordenadores actuales son tan rápidos que ahora la demostración completa puede verificarse en un ordenador casero en unas pocas horas.

Todo esto está muy bien, pero el ordenador sigue siendo el rey. ¿Podemos prescindir de él? Hay una sensación creciente de que en este caso concreto quizá no sea totalmente inconcebible una historia que los seres humanos pueden captar en su totalidad. Quizá nuevas ideas en el problema de los cuatro colores llevarán al fin a una demostración más sencilla, con poca o ninguna ayuda del

ordenador, de modo que los matemáticos puedan leerla, reflexionar sobre ella y decir «¡Sí!». Aún no conocemos esa demostración, y quizá no exista, pero hay una sensación en el aire…

Los matemáticos están aprendiendo mucho sobre redes. Topólogos y geómetras están encontrando profundas relaciones entre redes y áreas de las matemáticas completamente diferentes, incluidas algunas que se aplican a la física matemática. Uno de los conceptos que aparecen, de vez en cuando, es el de curvatura. El nombre es adecuado: la curvatura de un espacio nos dice cómo se deforma. Si es liso como el plano, su curvatura es cero. Si se curva en la misma dirección, de la misma manera que una colina se curva hacia abajo en todas partes, tiene curvatura positiva. Si es como un paso entre montañas, que se curva hacia arriba en unas direcciones pero hacia abajo en otras, tiene curvatura negativa. Hay teoremas geométricos, descendientes de la fórmula de Euler, que relacionan redes dibujadas en un espacio con la propia curvatura del espacio. A eso se refiere la fórmula de Heawood para un todo *g*-agujereado. Una esfera tiene curvatura positiva; un toro representado como un cuadrado con los lados opuestos identificados (véase Figura 12, *derecha*), tiene curvatura cero; y un toro con dos o más agujeros tiene curvatura negativa. De modo que hay cierto tipo de vínculo entre curvatura y coloreado de mapa.

Detrás de este vínculo hay una propiedad útil de la curvatura: es difícil deshacerse de ella. Es como un gato bajo una alfombra. Si la alfombra es plana, no hay gato, pero si se ve un bulto, hay un gato debajo. Se puede perseguir al gato por la alfombra, pero todo lo que

se consigue es mover el bulto de un lugar a otro. Del mismo modo, la curvatura puede moverse, pero no eliminarse. A menos que el gato llegue al borde de la alfombra, en cuyo caso puede escapar llevándose con él su curvatura. Las reglas de descarga de Heesch son un poco como la curvatura. Desplazan la carga eléctrica, pero no la destruyen. ¿Podría existir algún concepto de curvatura para una red, y algunas astutas reglas de descarga que, en efecto, desplacen la curvatura?

Si así fuera, uno podría ser capaz de persuadir a una red para colorearse de modo automático. Asignamos curvatura a sus puntos (y quizá líneas); luego dejamos que la red redistribuya la curvatura más uniformemente. Aquí «uniformemente» implica quizá que bastan cuatro colores si establecemos todo de forma correcta. Es solo una idea, no es mía, y no la he explicado con bastante detalle para que tenga mucho sentido. Pero refleja la intuición de algunos matemáticos, y ofrece esperanzas de que aún pueda encontrarse una demostración más conceptual del teorema de los cuatro colores —una historia entretenida antes que un resumen con mil millones de listines telefónicos como apéndice—. En el capítulo 10 encontraremos una idea similar, en un contexto mucho más sofisticado, que resolvió un problema aún mayor en topología.

Capítulo 5

Simetría esférica

La conjetura de Kepler

Todo empezó con un copo de nieve.

La nieve tiene una extraña belleza. Cae del cielo en suaves copos blancos, vuela en el viento para crear blandos montículos que cubren el paisaje, forma espontáneamente extrañas figuras. Es fría. Podemos esquiar sobre ella, viajar en un trineo, hacer bolas de nieve y muñecos de nieve con ella… y si no tenemos suerte, podemos quedar enterrados por miles de toneladas de ella. Cuando desaparece no vuelve directamente al cielo en forma de copos blancos. Se convierte en agua ordinaria. Puede evaporarse y volver al cielo, por supuesto, pero puede viajar por ríos hasta llegar al mar, y pasa mucho tiempo en los océanos. La nieve es una forma de hielo, y el hielo es agua congelada.

Esto no es nada nuevo. Quizá ya era obvio para los neandertales.

Los copos de nieve no son en absoluto grumos informes. Cuando son prístinos, antes de que empiecen a fundirse, muchos de ellos son estrellas minúsculas e intrincadas: planas, de seis lados y simétricas. Otros son simples hexágonos. Algunos tienen menos simetría, otros tienen una tercera dimensión apreciable, pero los copos de nieve con simetría séxtuple son icónicos y están muy extendidos. Los copos de nieve son cristales de hielo. Esto tampoco es nuevo: solo tenemos que reconocer un cristal cuando vemos uno. Pero no son cristales ordinarios, con caras planas y poligonales. Su

característica más enigmática añade un toque de caos: pese a tener la misma simetría, la estructura detallada difiere de un copo de nieve al siguiente. No hay dos copos de nieve iguales, dicen algunos. Siempre me he preguntado cómo lo saben, pero los números favorecen esa idea si se es suficientemente pedante sobre lo que cuenta como iguales.

¿Por qué los copos de nieve tienen seis lados? Hace cuatrocientos años, uno de los grandes matemáticos y astrónomos del siglo XVII se planteó esta pregunta, y sugirió una respuesta. Resultó ser una respuesta sorprendentemente buena, y más teniendo en cuenta que no hizo ningún experimento especial. Tan solo combinó algunas ideas simples que eran conocidas para todos. Por ejemplo, la forma en que las semillas de granada están empaquetadas dentro del fruto.

Su nombre era Johannes Kepler, y tenía una muy buena razón para pensar en los copos de nieve. Su sustento dependía de un rico mecenas, John Wacker de Wakenfels. En esa época Kepler era matemático de la corte de Rodolfo II, emperador del Sacro Imperio Romano, y Wacker, un diplomático, era consejero del emperador. Kepler quería dar a su mecenas un regalo de Año Nuevo. En principio, debería ser barato, inusual y estimulante. Wacker debería hacerse una idea de los notables descubrimientos que su dinero estaba haciendo posibles. De modo que Kepler reunió sus ideas sobre los copos de nieve en un librito, y ese fue el regalo. Su título era *De nive sexangula* («Sobre el copo de nieve de seis puntas»). Fue en 1611. Escondido en su interior, uno de los pasos importantes en

el pensamiento de Kepler, había un breve comentario: un enigma matemático que iba a tardar 387 años en resolverse.

Kepler era un empedernido buscador de pautas. Su trabajo científico más influyente fue el descubrimiento de tres leyes básicas del movimiento planetario, la primera y más conocida de las cuales es que la órbita es una elipse. También era un místico, totalmente imbuido de la idea platónica de que el universo se basa en números, pautas y formas matemáticas. Hacía astrología tanto como astronomía: en esa época los matemáticos solían hacer chapuzas como astrólogos, porque realmente podían hacer las cuentas para calcular cuando el ascendiente estaba en Acuario. Los patronos ricos, incluso la realeza, les pagaban por hacer horóscopos.

En su libro Kepler señalaba que la nieve empieza como vapor de agua, que es informe, pero de algún modo el vapor se convierte en copos sólidos de seis lados. Algún agente debía causar esa transición, insistía Kepler:

¿Imprime [este agente] la forma de seis puntas en la materia exigido por la materia, o lo hace por su propia naturaleza, una naturaleza, por ejemplo, en la que está innata bien la idea de la belleza inherente en el hexágono o bien el conocimiento del propósito que dicha forma favorece?

En busca de la respuesta, él consideró otros ejemplos de formas hexagonales en la naturaleza. En seguida vienen a la mente los panales de miel en las colmenas. Están hechos de dos capas de celdas hexagonales, adosadas, y sus extremos comunes están formados por tres rombos (paralelogramos con todos los lados

iguales). Esta forma le recordaba a Kepler la de un sólido llamado dodecaedro rómbico (véase Figura 15). No es uno de los cinco poliedros regulares que los pitagóricos conocían y Euclides clasificó, pero tiene una propiedad distintiva: pueden empaquetarse copias idénticas para llenar exactamente el espacio, sin dejar huecos. La misma forma se da en las granadas, donde pequeñas semillas redondas crecen, se aprietan y por consiguiente se ven forzadas a crear un empaquetamiento eficiente.

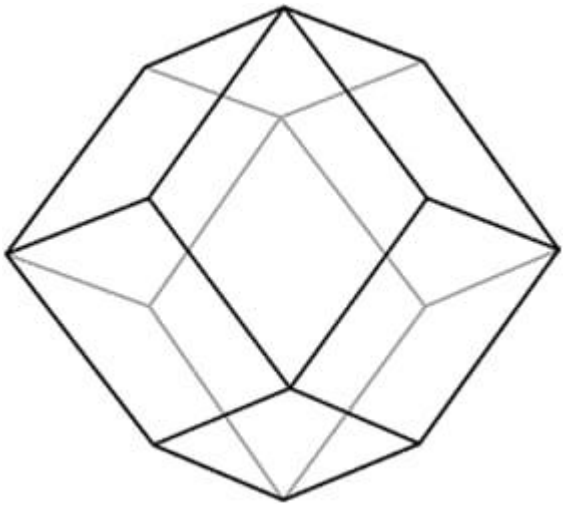


Figura 15. El dodecaedro rómbico, un sólido con 12 caras rómbicas.

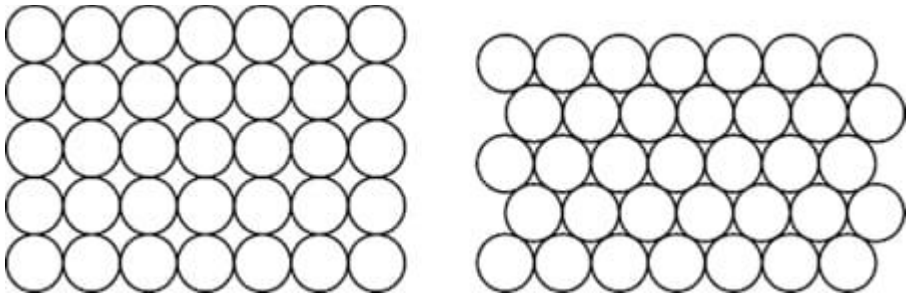


Figura 16. Izquierda: Empaquetamiento en red cuadrada. Derecha: Empaquetamiento en red triangular (también llamada hexagonal).

Como cualquier matemático razonable, Kepler empieza por el caso más simple en el que las esferas forman una única capa plana. Esto es equivalente a empaquetar círculos idénticos en el plano. Aquí encuentra solo dos disposiciones regulares. En una, las esferas se disponen en cuadrados (véase Figura 16, *izquierda*); en la otra se disponen en triángulos equiláteros (véase Figura 16, *derecha*). Estas disposiciones, repetidas a lo largo del plano infinito, son la red cuadrada y la red triangular. La palabra «red» hace referencia a esta pauta espacialmente periódica que se repite en dos direcciones independientes. Las figuras necesariamente muestran una porción finita de la pauta, así que deberíamos olvidarnos de los bordes. Lo mismo sucede con las Figuras 17-20 más abajo. La Figura 16, izquierda y derecha, muestran cinco filas de esferas, y en cada fila las esferas tocan a sus vecinas. Sin embargo, la red triangular está ligeramente comprimida: sus filas están más juntas. Por ello las esferas en la red triangular están más estrechamente empaquetadas que las de la red cuadrada.

A continuación, Kepler se pregunta cómo pueden colocarse capas sucesivas de este tipo unas encima de otras, y considera cuatro casos. En los dos primeros, todas las capas son redes cuadradas. Una forma de apilar las capas es colocar las esferas de cada capa directamente encima de las que están debajo. Entonces cada esfera tendrá seis primeras vecinas: cuatro dentro de su capa, una encima y una debajo. Este empaquetamiento es como un tablero de ajedrez tridimensional hecho de cubos, y en eso se convertiría si se inflaran las esferas hasta que ya no pudieran expandirse más. Pero esto,

dice Kepler, «no será el empaquetamiento más apretado». Puede apretarse más deslizando lateralmente la segunda capa, de modo que sus esferas encajen limpiamente en las hendiduras entre las esferas de la capa inferior (véase Figura 17, *izquierda*). Repitamos este proceso, capa a capa (Figura 17, *derecha*). Ahora cada esfera tiene doce vecinas: cuatro en su propia capa, cuatro arriba y cuatro abajo. Si las inflamos el espacio se llenará con dodecaedros rómbicos.

En los otros dos casos, las capas son redes triangulares. Si están apiladas de modo que las esferas de cada capa caen directamente sobre las que están debajo, entonces cada esfera tiene ocho vecinas: seis en su propia capa, una encima y una debajo. Alternativamente, las esferas en la capa siguiente pueden estar de nuevo encajadas en las hendiduras de la capa inferior. Ahora cada esfera tiene doce vecinas: seis en su propia capa, tres encima y tres debajo. Este es el mismo número de vecinas que las esferas en la segunda disposición de capas cuadradas, y Kepler ofrece un cuidadoso análisis de la geometría para mostrar que esta cuarta disposición es en realidad la misma que la segunda. La única diferencia es que las capas cuadradas ya no son horizontales sino que están inclinadas un ángulo. Él escribe: «Así pues, el empaquetamiento más apretado en tres dimensiones, la pauta triangular, no puede existir sin la cuadrada, y viceversa». Volveré a esto: es importante.

Tras exponer la geometría básica del empaquetamiento de esferas, Kepler vuelve al copo de nieve y la simetría séxtuple. Recuerda el empaquetamiento en red triangular de esferas en un plano, en el

que cada esfera está rodeada por otras seis, que forman un hexágono perfecto. Esta, decide él, debe ser la causa de que los copos de nieve tengan seis lados.

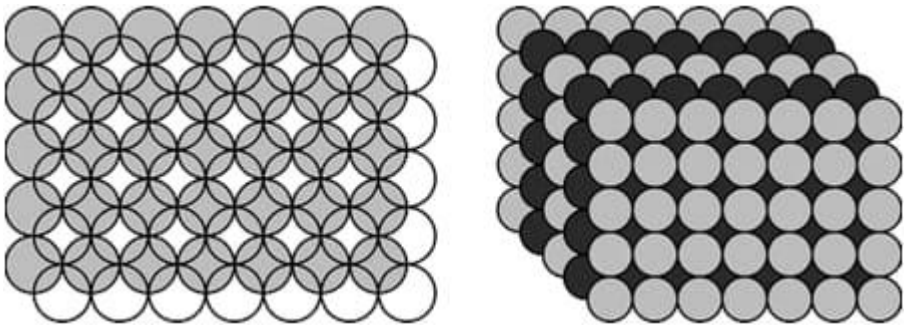


Figura 17. Izquierda: Añadiendo una segunda capa de esferas (círculos abiertos) sobre la primera capa (gris). Derecha: Repitiendo esta construcción.

Este capítulo no trata principalmente de copos de nieve, pero la explicación de Kepler de su simetría es muy similar a la que daríamos hoy, de modo que sería una lástima detenernos aquí. ¿Por qué son —cómo pueden ser— tan variados, y pese a todo simétricos? Cuando el agua cristaliza para formar hielo los átomos de hidrógeno y oxígeno que forman las moléculas de agua se empaquetan en una estructura simétrica, la red cristalina. Esta red es más complicada que cualquiera de las disposiciones de esferas de Kepler, pero su simetría dominante es séxtuple. Un copo de nieve crece a partir de una minúscula «semilla» de solo unos pocos átomos dispuestos como un pequeño fragmento de red. Esta semilla tiene la misma simetría séxtuple y fija el escenario para el

crecimiento del cristal de hielo cuando los vientos lo zarandean de aquí para allá dentro de una nube de tormenta.

La gran variedad de pautas de copos de nieve es una consecuencia de las condiciones variables en la nube. Dependiendo de cuáles sean la temperatura y la humedad, el crecimiento cristalino puede ser uniforme o puede ser dendrítico. En el primer caso, los átomos se van añadiendo al mismo ritmo a lo largo de todo el contorno, lo que lleva a hexágonos de lados rectos. En el segundo, el ritmo de crecimiento varía de un lugar a otro, lo que da lugar a estructuras arborescentes. Conforme el copo en crecimiento es movido arriba y abajo a través de la nube, estas condiciones siguen cambiando, aleatoriamente. Pero el copo es tan minúsculo que en cualquier momento dado las condiciones son en esencia las mismas en las seis puntas, de modo que todas hacen lo mismo. Todo copo de nieve lleva trazas de su historia. En la práctica, la simetría séxtuple nunca es exacta, pero suele estar muy próxima. El hielo es una materia extraña, y también son posibles otras formas: puntas, placas planas, prismas hexagonales, prismas con placas en sus extremos. La historia completa es muy complicada, pero todo se basa en cómo están dispuestos los átomos en los cristales de hielo[38]. En los días de Kepler la teoría atómica era, como mucho, una vaga sugerencia de algunos antiguos griegos; es sorprendente hasta dónde llegó sobre la base de observaciones rústicas, experimentos mentales y una sensibilidad por las pautas.

[38] Una página excelente para información sobre la física de los copos de nieve es http://www.its.caltech.edu/-atomic/snowcrystals/.

La conjetura de Kepler no se refiere a los copos de nieve como tales. La conjetura es su comentario displicente de que apilar capas de esferas estrechamente empaquetadas, de modo que capas sucesivas encajen en las hendiduras que quedan entre las de la capa anterior, lleva «al empaquetamiento más apretado en tres dimensiones». La conjetura puede resumirse de manera informal: si se quiere empaquetar muchas naranjas en una caja grande, ocupando el máximo espacio posible, habría que empaquetarlas como lo haría cualquier frutero.

La dificultad no está en encontrar la *respuesta*. Kepler nos la dio. Lo difícil es demostrar que tenía razón. Durante siglos se acumularon muchas pruebas indirectas. Nadie pudo dar con un empaquetamiento más estrecho. La misma disposición de átomos es común en cristales, donde el empaquetamiento eficiente corresponde presumiblemente a minimizar la energía, un principio estándar que gobierna muchas formas naturales. Este tipo de pruebas eran bastante buenas para satisfacer a la mayoría de los físicos. Por otra parte, nadie podía dar con una demostración de que *no había* nada mejor. Cuestiones más sencillas del mismo tipo, tales como empaquetar círculos en el plano, resultaron tener profundidades ocultas. El área entera era difícil y estaba llena de sorpresas. Todo esto preocupaba a los matemáticos, incluso si la mayoría de ellos pensaba que Kepler tenía la respuesta correcta. En 1958 C. Ambrose Rogers describió la conjetura de Kepler como algo

que «muchos matemáticos creen, y todos los físicos saben[39]». Este capítulo describe cómo los matemáticos convirtieron esta creencia en certeza.

Para entender lo que hicieron tenemos que echar una mirada más cercana a la disposición de esferas de Kepler, que es conocida como la red cúbica de caras centradas. Cuando lo hacemos empiezan a manifestarse las sutilezas del problema. La primera cuestión que viene a la mente es por qué utilizamos redes cuadradas. Después de todo, el empaquetamiento más apretado en una única capa se da para la red triangular. La respuesta es que también podemos obtener la red cúbica de caras centradas utilizando capas triangulares; esta es la esencia del comentario de Kepler de que «la pauta triangular no puede existir sin la cuadrada». Sin embargo, es más fácil describir la red cúbica de caras centradas utilizando capas cuadradas. Como premio añadido vemos que la conjetura de Kepler no es tan sencilla como el empaquetamiento de naranjas por los fruteros.

Supongamos que empezamos con una capa plana de esferas dispuestas en triángulos (véase Figura 16, *derecha*). Entre las esferas hay huecos triangulares curvos, y otra capa de esferas puede encajar en estos. Cuando empezábamos con una red cuadrada éramos capaces de utilizar todos los huecos, de modo que la posición de la segunda capa, y las que seguían, estaba unívocamente determinada. Este ya no es el caso si empezamos con una disposición triangular. No podemos utilizar todos los huecos

[39] C. A. Rogers, «The packing of equal spheres», *Proceedings of the London Mathematical Society*, n.º 8 (1958), pp. 609-620.

porque están demasiado juntos. Solo podemos utilizar la mitad. Una elección se muestra en la Figura 18 (*izquierda*), utilizando pequeños puntos grises por claridad, y la Figura 18 (*derecha*) muestra cómo debería colocarse la siguiente capa de esferas. La segunda manera de encajar una nueva capa en los huecos de la capa 1 se muestra en la Figura 19 (*izquierda*) utilizando puntos más oscuros. Estos puntos coinciden con huecos en la capa 2, de modo que añadimos la capa 3 en las posiciones correspondientes: el resultado es la Figura 19 (*derecha*).

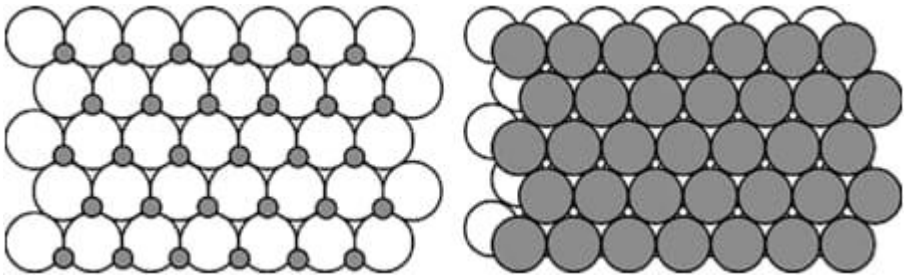


Figura 18. Encajando una red triangular en un conjunto de huecos en la cara inferior.

La distinción entre estas elecciones no supone en realidad ninguna diferencia cuando solo tenemos estas dos capas. Si rotamos 60 ° la segunda disposición, obtenemos la primera. Son iguales «salvo simetría». Pero una vez que han sido colocadas las dos primeras capas, hay dos elecciones genuinamente diferentes para la tercera. Cada nueva capa tiene dos sistemas de huecos, mostrados por los puntos claros y oscuros en la Figura 19 (*izquierda*). Un sistema coincide con los centros de la capa inmediata de debajo, visibles como pequeños triángulos gris claro en la Figura 19 (*derecha*). El

otro coincide con los huecos en la capa por debajo de esta, visibles como triángulos que contienen un minúsculo hexágono blanco en la Figura 19 (*derecha*). Para obtener la red cúbica de caras centradas debemos utilizar las posiciones gris oscuro para la tercera capa, y luego continuar la misma pauta indefinidamente.

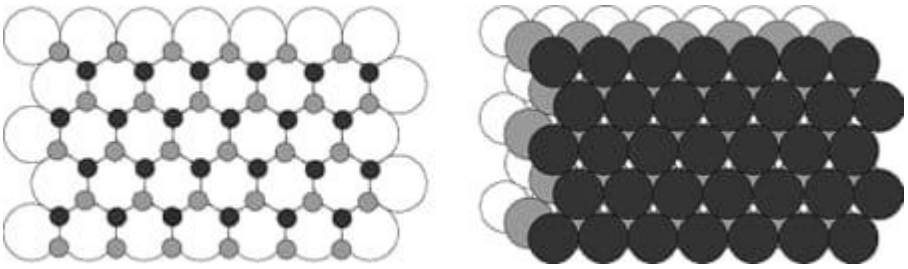


Figura 19. Apilar unas redes triangulares unas encima de otras.

No es del todo obvio que el resultado sea la red cúbica centrada en las caras. ¿Dónde están los cuadrados? La respuesta es que están presentes, pero inclinados cierto ángulo. La Figura 20 muestra seis capas triangulares sucesivas, con algunas esferas eliminadas. Las flechas indican las filas y columnas de una red cuadrada, oculta en el interior. Capas paralelas a esta son también redes cuadradas, y encajan exactamente en la manera en que he construido la red cúbica de caras centradas.

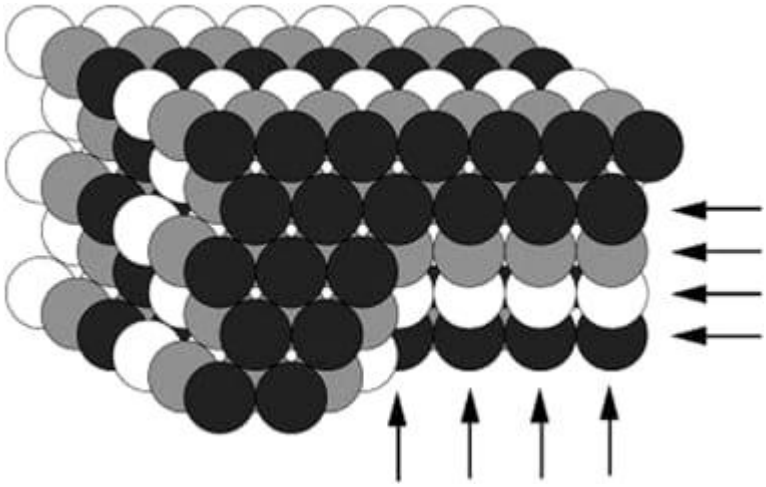


Figura 20. Ocultas dentro de las capas triangulares hay capas cuadradas, con una inclinación.

¿Cuán «estrecho» es este empaquetamiento? Medimos la estrechez (eficiencia, proximidad) de un empaquetamiento por su densidad: la proporción de espacio ocupada por las esferas[40]. Cuanto mayor es la densidad, más estrecho es el empaquetamiento. Los cubos se empaquetan con densidad 1, llenando todo el espacio. Como es obvio, las esferas tienen que dejar huecos, de modo que la densidad es menor que 1. En el caso de la red cúbica de caras centradas la densidad es $\pi/\sqrt{18}$, que es aproximadamente 0,7405. De modo que en este empaquetamiento las esferas llenan algo menos que las tres cuartas partes del espacio. La conjetura de Kepler afirma que ningún empaquetamiento de esferas puede tener una densidad mayor que esta.

[40] Puesto que el espacio es infinito, hay infinitas esferas, de modo que el espacio y las esferas tienen un volumen total infinito. No podemos definir la densidad como ∞/∞, porque eso no tiene un valor numérico bien definido. En su lugar consideramos regiones de espacio cada vez más grandes y tomamos el valor límite de la fracción de dichas regiones que llenan las esferas.

Lo he afirmado con mucho cuidado. No he dicho «el empaquetamiento cúbico de caras centradas tiene una densidad mayor que cualquier otro». Eso es falso, espectacularmente falso. Para ver por qué, volvamos a la construcción de la red cúbica de caras centradas utilizando capas triangulares. Dije que una vez que se han colocado las dos primeras capas, hay dos elecciones para la tercera capa. La red cúbica de caras centradas aparece si utilizamos la segunda, los puntos gris oscuro. ¿Qué sucede si utilizamos la otra, los puntos gris claro? Ahora la capa 3 se sitúa exactamente sobre la capa 1. Si continuamos así, colocando cada nueva capa de modo inmediato sobre la capa dos etapas por debajo, obtenemos una segunda red de empaquetamiento: la red hexagonal. Es por completo diferente del empaquetamiento cúbico de caras centradas, pero tiene la misma densidad. Esto es obvio porque las dos maneras diferentes de colocar la tercera capa están relacionadas por una simetría rotacional, de modo que su encaje con la capa anterior es igualmente estrecho de las dos maneras.

Estos son los dos únicos empaquetamientos en *red* que se pueden obtener a partir de sucesivas capas triangulares, pero en 1883 el geólogo y cristalógrafo William Barlow señaló que podemos escoger al azar entre las dos posibilidades la localización de cada capa sucesiva. Puesto que una u otra posición dan la misma contribución a la densidad, todos estos empaquetamientos tienen densidad $\pi/\sqrt{18}$. Hay infinitas secuencias aleatorias, lo que lleva a infinitos empaquetamientos diferentes, todos con la misma densidad.

En resumen, no hay tal cosa como «el» empaquetamiento de esferas más denso. Más bien, hay infinitos de ellos, todos igual de densos. Esta falta de unicidad es una advertencia de que esto no es un problema simple. La *densidad* óptima es única, si Kepler tenía razón, pero existen infinitas disposiciones diferentes con esa misma densidad. De modo que una demostración de que esta densidad realmente es óptima no es solo cuestión de encajar sucesivamente cada nueva esfera tan prieta como sea posible. Hay elecciones.

Por impresionante que pueda ser la experiencia de los fruteros —y la red cúbica de caras centradas estaba sin duda presente en los mercados del Egipto predinástico— no es ni mucho menos concluyente. De hecho, que el método del frutero dé una buena respuesta es algo accidental. El problema al que se enfrentan los fruteros no es empaquetar naranjas tan estrechamente como se pueda en el espacio, donde cualquier disposición es posible en principio. Su problema es apilar naranjas de modo estable, en un mundo donde el suelo es plano y la gravedad actúa hacia abajo. Los fruteros empiezan, como es natural, formando una capa; luego añaden otra capa, y así sucesivamente. Sin duda hacen con la primera capa una red cuadrada si están colocando las naranjas dentro de una caja rectangular. Si las naranjas no están confinadas, entonces una red cuadrada o una red triangular es lo natural. En la práctica, ambas dan la misma red cúbica centrada en las caras (al menos, si las capas se colocan de modo adecuado en el caso triangular). La red cuadrada parece realmente una pobre elección,

porque no es la forma más densa de empaquetar una capa. Por suerte, más que por buen juicio, resulta que eso no importa.

Los físicos no están interesados en las naranjas. Lo que ellos quieren empaquetar son átomos. Un cristal es una disposición de átomos regular y espacialmente periódica. La conjetura de Kepler explica la periodicidad como una consecuencia natural de que los átomos se empaquetan del modo más estrecho posible. Para los físicos, la existencia de cristales es prueba suficiente, así que la conjetura es evidentemente cierta. Sin embargo, acabamos de ver que hay infinitas maneras de empaquetar esferas tan densamente como lo hacen la red cúbica de caras centradas y la red hexagonal, pero ninguna ellas es espacialmente periódica. Entonces, ¿por qué la naturaleza utiliza pautas periódicas para los cristales? Una posible respuesta es que no deberíamos modelar los átomos como esferas.

Tampoco los matemáticos están interesados en las naranjas. Como Kepler, ellos prefieren trabajar con esferas perfectas e idénticas. No encuentran convincente el argumento de los físicos. Si no debiéramos modelar los átomos como esferas, la existencia de cristales deja de ser una prueba a favor de la conjetura de Kepler. No podemos tenerla de las dos maneras. Incluso si se argumenta que la conjetura explica en parte la red cristalina y la red cristalina muestra en parte que la conjetura es correcta… hay una laguna lógica. Los matemáticos quieren una demostración.

Kepler no llamó conjetura a su afirmación: simplemente la puso en su libro. No está nada claro si él pretendía que se interpretara de

una manera tan amplia. ¿Estaba afirmando que la red cúbica de caras centradas era el «empaquetamiento más estrecho en tres dimensiones» entre todas las maneras concebibles de empaquetar esferas? ¿O simplemente quería decir que era el empaquetamiento más estrecho de los tres que había considerado? No podemos retroceder al pasado y preguntar. Cualquiera que sea la realidad histórica, la interpretación de interés para los matemáticos y físicos era la amplia, la ambiciosa. La que pedía contemplar toda manera posible de empaquetar infinitas esferas en un espacio infinito, y mostrar que ninguna de ellas tiene densidad mayor que la red cúbica de caras centradas.

Es muy fácil subestimar la dificultad de la conjetura de Kepler. Sin duda, la forma de obtener el empaquetamiento más apretado es añadir esferas una a una, haciendo que cada una toque a tantas otras como sea posible. Esto lleva inevitablemente a la pauta de Kepler. Y así sucede si se añaden las esferas en el orden correcto, colocándolas en las posiciones correctas cuando hay alternativas. Sin embargo, no hay garantía de que este proceso paso a paso, añadiendo esferas de una en una, no puede ser superado por algo de mayor alcance. Cualquiera que haya colocado el equipaje de vacaciones en el maletero de un automóvil aprende que encajar los bultos de uno en uno puede dejar huecos en los que nada encaja, pero volviendo a empezar y teniendo más cuidado, a veces cabe más. Por supuesto, parte del problema de empaquetar el equipaje de vacaciones está en las diferentes formas y tamaños de los objetos que se trata de encajar, pero el punto lógico es bastante claro:

asegurar la disposición más apretada en una región pequeña podría tener repercusiones y no llevar a la disposición más apretada en una región más grande.

Las disposiciones que considera Kepler son muy especiales. Es concebible que una disposición por completo diferente pudiera empaquetar esferas idénticas más apretadamente. Quizá capas con baches fueran más eficientes. Quizá «capas» es la idea equivocada. E incluso si uno está absolutamente seguro de que es la idea correcta, aún hay que demostrarlo.

¿No está convencido? ¿Sigue pensando que es obvio? ¿Tan obvio que no necesita demostración? Permítame destruir su confianza en su intuición sobre el empaquetamiento de esferas. He aquí una pregunta mucho más sencilla concerniente a círculos en el plano. Suponga que yo le doy 49 círculos idénticos, cada uno de diámetro unidad. ¿Cuál es el tamaño del cuadrado *más pequeño* que los contiene, si están empaquetados sin solaparse? La Figura 21 (*izquierda*) muestra la respuesta obvia: empaquetarlos como botellas de leche en una caja. El lado de la caja es exactamente 7 unidades. Para probar que esto es lo mejor, observe que cada círculo es mantenido fijo en su lugar por todos los demás, así que no hay modo de crear espacio extra. La Figura 21 (*derecha*) muestra que esta respuesta es errónea. Empaquetémoslos de la manera irregular mostrada y encajarán en una caja cuadrada cuyo lado es ligeramente menor que 6,98 unidades[41]. De modo que la

[41] http://hydra.nat.uni-magdeburg.de/packing/csq/csq49.html.

demostración también es errónea. Ser riguroso no es garantía de que no se puede hacer mejor.

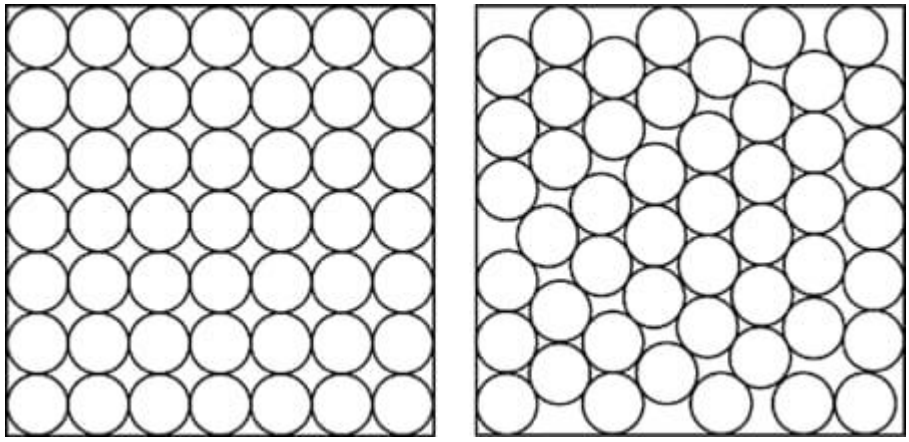


Figura 21. Izquierda: 49 círculos en un cuadrado de 7×7. Derecha: Cómo encajar 49 círculos en un cuadrado ligeramente menor.

De hecho, es fácil ver que el razonamiento que lleva a la respuesta «7» no puede ser correcto. Tan solo consideremos cuadrados más grandes. Utilizando una red cuadrada, n^2 círculos de diámetro 1 se empaquetan en un cuadrado de lado n. No hay manera de mejorar la densidad moviendo estos círculos de forma continua, porque el empaquetamiento es rígido. Pero debe haber empaquetamientos más densos para n suficientemente grande, porque una red triangular es más eficiente que una red cuadrada. Si tomamos un cuadrado realmente grande, y encajamos tantos círculos en él como podamos utilizando una red triangular, la ventaja que esto tiene sobre la red cuadrada hará que al final gane, pese a «efectos de borde» en el contorno donde hay que dejar huecos. El tamaño del contorno es $4n$, que se hace arbitrariamente pequeño comparado

con n^2. Resulta que el punto exacto en el que domina la red triangular es cuando $n = 7$. Esto no es obvio y se necesita mucho trabajo detallado para establecerlo, pero algún n tiene que funcionar. La rigidez no es suficiente.

En realidad hay dos versiones de la conjetura de Kepler. Una considera solo el empaquetamientos en red, donde los centros de las esferas forman una pauta espacialmente periódica que se repite de modo indefinido en tres direcciones independientes como una especie de papel de pared sólido. Incluso así, el problema es difícil, porque hay muchas redes diferentes en el espacio. Los cristalógrafos reconocen catorce tipos, clasificados por sus simetrías, y algunos de estos tipos están determinados por números que pueden ajustarse a infinitos valores diferentes. Pero las dificultades se agravan cuando consideramos la segunda versión del problema, que permite todos los empaquetamientos posibles. Cada esfera se cierne en el espacio, no hay gravedad y no hay obligación de formar capas u otras disposiciones simétricas.

Cuando un problema parece demasiado difícil, los matemáticos lo dejan en segundo plano y buscan versiones más sencillas. Las ideas de Kepler sobre capas planas de esferas sugieren empezar con empaquetamientos de círculos en un plano. Es decir, dado un suministro infinito de círculos idénticos, empaquetarlos lo más estrechamente posible. Ahora la densidad es la proporción de área que cubren los círculos. En 1773, Joseph Louis Lagrange demostró que el empaquetamiento en red más denso de círculos en un plano es la red triangular, con densidad $\pi/\sqrt{12} = 0{,}9069$. En 1831 Gauss

estaba revisando un libro de Ludwig Seeber, quien había generalizado algunos de los resultados de Gauss en teoría de números a ecuaciones en tres variables. Gauss comentó que los resultados de Seeber demuestran que las redes hexagonal y cúbica de caras centradas proporcionan el empaquetamiento más denso en el espacio tridimensional. Mucho se sabe ahora sobre redes de empaquetamiento en espacios de dimensión más alta —4, 5, 6 y así sucesivamente—. El caso 24-dimensional está especialmente bien entendido. (Así son las cosas). Pese a su aire de imposibilidad práctica, esta área tiene en realidad implicaciones para la teoría de la información y los códigos de computación.

Los empaquetamientos no reticulares son otra cuestión completamente diferente. Existen infinitos, y no tienen ninguna bonita estructura regular. Entonces ¿por qué no ir al otro extremo y ensayar empaquetamientos aleatorios? En su *Vegetable Staticks* de 1727, Stephen Hales informaba de experimentos en los que «comprimía varios lotes frescos de guisantes en el mismo puchero» y encontraba que, cuando todos se presionaban, formaban «bonitos dodecaedros regulares». Parece que quería decir que los dodecaedros regulares eran bellos, no que los dodecaedros fueran muy regulares, pero la segunda interpretación es mejor porque los dodecaedros regulares no pueden llenar el espacio. Lo que él vio eran sin duda dodecaedros rómbicos, que ya hemos visto que están asociados con el empaquetamiento cúbico de caras centradas. G. David Scott puso montones de bolas de cojinete en un recipiente y lo agitó violentamente, y observó que la densidad más alta era 0,6366. En

2008 Chaoming Song, Ping Wang y Hernán Makse dedujeron está cifra analíticamente[42]. Sin embargo, su resultado no implica que Kepler tuviera razón, aunque solo sea porque, tal como está establecido, implicaría que la red cúbica de caras centradas, con densidad 0,74, no puede existir. La manera más sencilla de explicar esta discrepancia es que su resultado ignora excepciones extremadamente raras. La red cúbica de caras centradas, la red hexagonal y todas las disposiciones de capas triangulares arbitrariamente escogidas son excepciones de este tipo. Por la misma razón, podrían existir algunas otras disposiciones con una densidad aún mayor. No puede ser una red, pero una búsqueda aleatoria nunca la encontrará porque su probabilidad es cero. De modo que el estudio de empaquetamientos aleatorios, aunque relevante para muchas cuestiones en física, no nos dice mucho sobre la conjetura de Kepler.

El primer gran avance real llegó en 1892, cuando Axel Thue dio una conferencia en el Congreso Escandinavo de Ciencia Natural en la que esbozó una demostración de que ningún empaquetamiento de círculos en un plano puede ser más denso que la red triangular. Su conferencia fue publicada, pero los detalles son demasiado vagos para reconstruir la demostración que tenía en mente. Dio una nueva demostración en 1910, que parecía convincente salvo por unos pocos puntos técnicos que él sencillamente supuso que podían resolverse. En lugar de llenar estas lagunas, László Fejes Tóth obtuvo una demostración completa por otros métodos en 1940. Poco

[42] C. Song, P. Wang y H. A. Makse, A phase diagram for jammed matter, *Nature*, n.º 453 (29 de mayo de 2008), pp. 629-632.

después, Beniamino Segre y Kurt Mahler encontraron demostraciones alternativas. En 2010 Hai-Chau Chang y Lih-Chung Wang pusieron en la web una demostración más sencilla[43].

Encontrar la máxima densidad para empaquetamientos de círculos o esferas, bajo condiciones especificadas, cae en una clase general de cuestiones matemáticas conocida como problemas de optimización. Un problema semejante busca el máximo o el mínimo valor de cierta función. Una función es una regla matemática para calcular una cantidad que depende de una forma específica de cierto conjunto de variables. La regla suele estar especificada por una fórmula, pero esto no es esencial. Por ejemplo, el problema de la caja de leche, con 49 círculos, puede formularse de esta manera. Las variables son las coordenadas de los centros de los 49 círculos; puesto que cada círculo necesita dos coordenadas, hay 98 variables. La función es el tamaño del cuadrado más pequeño, con lados paralelos a los ejes de coordenadas, que contiene un conjunto dado de círculos que no se solapan El problema de la caja de leche es equivalente a encontrar el valor mínimo que puede alcanzar esta función cuando las variables recorren todos los empaquetamientos.

[43] Hai-Chau Chang y Lih-Chung Wang, «A simple proof of Thue's theoreme on circle packing», arXiv:1009.4322v1 (2010).

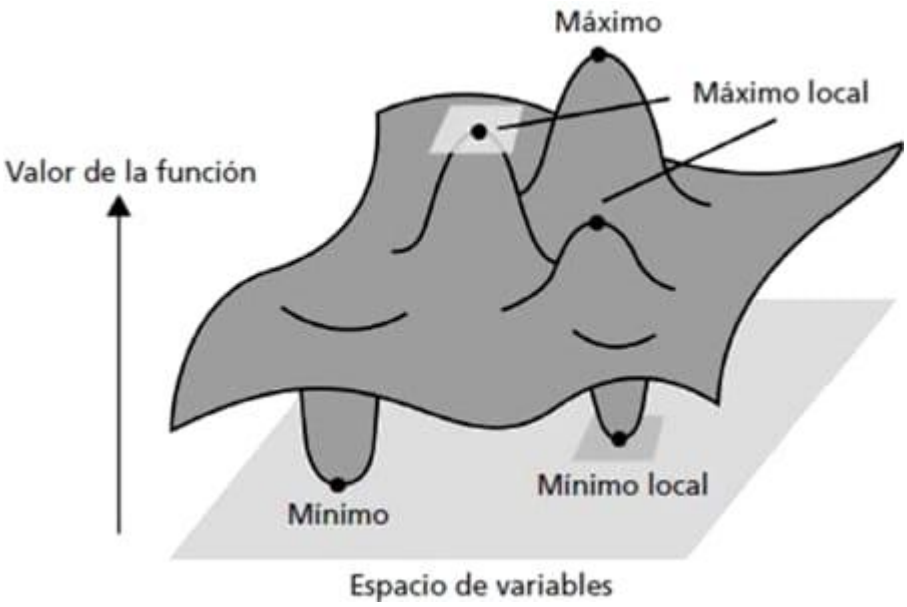


Figura 22. Picos y valles de una función.

Una función puede considerarse como un paisaje multidimensional. Cada punto en el paisaje corresponde a una elección de las variables, y la altura en dicho punto es el correspondiente valor de la función. El máximo de la función es la altura del pico más alto, y el mínimo es la profundidad del valle más profundo. En principio, los problemas de optimización pueden resolverse mediante el cálculo infinitesimal: la función debe ser horizontal en un pico o un valle (véase Figura 22), y el cálculo infinitesimal expresa esta condición como una ecuación. Para resolver el problema de la caja de leche por este método tendríamos que resolver un sistema de 98 ecuaciones con 98 variables.

Una dificultad con los problemas de optimización es que ecuaciones como estas suelen tener muchas soluciones. Un paisaje puede tener muchos picos locales, y solo uno de ellos es el más alto. Pensemos en el Himalaya: casi todo son picos, *pero* solo el Everest tiene el

récord de altura. Los métodos para encontrar picos, de los que el más obvio es «ir cuesta arriba si se puede», quedan a menudo atrapados en un pico local. Otra dificultad es que a medida que crece el número de variables, también lo hace el número probable de picos locales. No obstante, este método funciona a veces. Incluso resultados parciales pueden ser útiles: si se encuentra un pico local, el máximo debe ser al menos de esa altura. Así es como se encontró la disposición mejorada de círculos en el problema de la caja de leche.

Para empaquetamientos en red, la función cuyo máximo se busca depende solo de un número finito de variables, las direcciones y longitudes a lo largo de las que se repite la red. Para empaquetamientos no reticulares, la función depende de infinitas variables: los centros de todos los círculos o esferas. En tales casos la utilización directa del cálculo infinitesimal u otras técnicas de optimización es imposible. La demostración de Tóth utilizaba una idea ingeniosa para reformular el problema del empaquetamiento no reticular de círculos como un problema de optimización en un conjunto de variables finito. Más tarde, en 1953, él se dio cuenta de que el mismo truco podía aplicarse en principio a la conjetura de Kepler. Por desgracia, la función resultante depende de unas ciento cincuenta variables, demasiadas para un cálculo a mano. Pero Tóth preveía una posible salida: «Teniendo en cuenta el rápido desarrollo de nuestros ordenadores, es imaginable que pueda determinarse el mínimo con gran exactitud».

En ese momento la computación estaba en su infancia y no existía ninguna máquina con la potencia suficiente. Por ello, los progresos posteriores en la conjetura de Kepler siguieron caminos diferentes. Varios matemáticos pusieron cotas —límites superiores— a cuán denso podía ser un empaquetamiento de esferas. Por ejemplo, en 1958 Rogers demostró que como máximo es 0,7797: no había ninguna excepción, esta cota se aplicaba a todos los empaquetamientos de esferas. En 1986, J. H. Lindsey mejoró la cota a 0,77844, y Douglas Muder la redujo un poco más en 1988 para obtener una cota de 0,77836[44]. Estos resultados muestran que no se puede hacer mucho *mejor* que el valor 0,7405 de la red cúbica centrada en las caras. Pero todavía quedaba una laguna, y pocas perspectivas de deshacerse de ella.

En 1990 Wu-Yi Hsiang, un matemático norteamericano, anunció una demostración de la conjetura de Kepler. Sin embargo, cuando se hicieron públicos los detalles no tardaron en surgir dudas. Cuando Tóth revisó el artículo en *Mathematical Reviews*, escribió: «Si se me pregunta [si el artículo proporciona] una demostración de la conjetura de Kepler, mi respuesta es no. Confío en que Hsiang llenará los detalles, pero creo que la mayor parte del trabajo está aún por hacer».

Thomas Hales, quien había estado trabajando en la conjetura durante muchos años, también dudaba de que el método de Hsiang pudiera ser reparado. En su lugar, él decidió que era el momento de tomar en serio la aproximación de Tóth. Había crecido una nueva

[44] J. H. Lindsey, «Sphere packing in R^3», *Mathematika,* n.º 33 (1986), pp. 137-147.

generación de matemáticos para quienes recurrir a un ordenador era más natural que recurrir a una tabla de logaritmos. En 1996 Hales esbozó una estrategia de demostración basada en la idea de Tóth. Requería identificar todas las maneras posibles de disponer varias esferas en la inmediata vecindad de una dada. Un empaquetamiento de esferas está determinado por los centros de las esferas; para esferas unidad, estos deben estar alejados al menos 2 unidades. Digamos que dos esferas son *vecinas* si sus centros están alejados a lo sumo 2,51 unidades. Este valor es cuestión de juicio: hagámoslo demasiado pequeño y no hay espacio suficiente para reordenar vecinos para mejorar la densidad; hagámoslo demasiado grande y el número de maneras de reordenar los vecinos se hace gigantesco. Hales encontró que 2,51 era un compromiso efectivo. Ahora podemos representar cómo se disponen los vecinos formando una red infinita en el espacio. Sus puntos son los centros de las esferas, y dos puntos están unidos por una línea si son vecinos. Esta red es una especie de esqueleto del empaquetamiento y contiene información vital sobre la vecindad de cada esfera.

Dada una esfera cualquiera, podemos examinar sus vecinos inmediatos en la red y considerar solo las líneas entre estos vecinos, omitiendo la esfera original. El resultado es una especie de jaula que rodea al punto en el centro de la esfera original. La Figura 23 (par izquierdo) muestra los vecinos de una esfera en la red cúbica de caras centradas y la jaula asociada. La Figura 23 (par derecho) hace lo mismo para una disposición de esferas especial, el prisma pentagonal, que resultó ser un actor clave en la demostración. Aquí

hay dos bandas de pentágonos paralelos al «ecuador» de la esfera central, más una única esfera en cada polo.

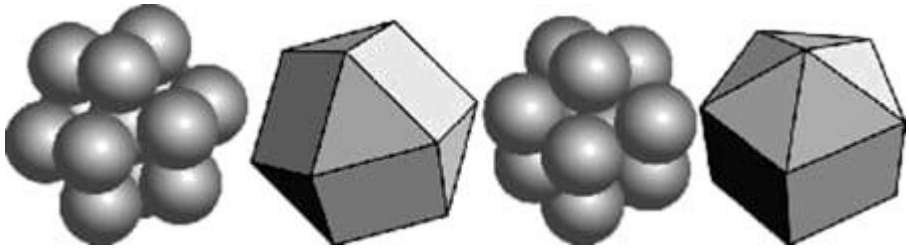


Figura 23. De izquierda a derecha: Vecindad de una esfera en la red cúbica de caras centradas; la jaula formada por sus vecinos; vecindad de una esfera de tipo prisma pentagonal; la jaula formada por sus vecinos.

Las jaulas forman un sólido con caras planas, y la geometría de este sólido controla la densidad de empaquetamiento cerca de la esfera central[45]. La idea clave consiste en asociar a cada jaula un número, conocido como su puntuación, que puede considerarse como una manera de estimar la densidad con la que están empaquetados los vecinos de la esfera. La puntuación no es la densidad propiamente dicha sino una cantidad que se comporta mejor y es más fácil de calcular. En particular, se puede encontrar la puntuación de la jaula sumando puntuaciones relacionadas con sus caras, lo que no funciona para la densidad. En general muchas nociones diferentes de puntuación satisfacen esta condición, pero todas coinciden en una cosa: para las redes hexagonal y cúbica de caras centradas la

[45] Hales utilizó varias nociones diferentes para lo que yo estoy llamando un jaula. La última es «estrella en descomposición». Mi descripción omite algunas distinciones cruciales para hacer comprensible la idea básica.

puntuación es siempre 8 «puntos», independientemente de qué elecciones se hagan en su definición. Aquí un punto es un número específico:

$$4\tan^{-1}\frac{\sqrt{2}}{5}-\frac{\pi}{3}=0{,}0553736$$

Así pues, 8 puntos es realmente 0,4429888. Este curioso número procede de la geometría especial de la red cúbica de caras centradas. La observación clave de Hales relaciona la conjetura de Kepler con este número: si toda jaula tiene una puntuación de 8 puntos o menos, entonces la conjetura de Kepler es cierta. De modo que el foco se desplaza a las jaulas y puntuaciones.

Las jaulas pueden clasificarse por su topología: cuántas caras tienen con un número dado de lados y cómo se unen estas caras. Sin embargo, para una topología dada los lados pueden tener muchas longitudes diferentes. Las longitudes afectan a la puntuación, pero la topología agrupa montones de jaulas diferentes y estas pueden tratarse de la misma forma general. En su demostración final, Hales consideró unos cinco mil tipos de jaula, pero los cálculos principales se centraban en unos pocos centenares. En 1992, él propuso un programa de cinco etapas:

1. Demostrar el resultado deseado cuando todas las caras de la jaula son triángulos.

2. Mostrar que los empaquetamientos hexagonal y cúbico de caras centradas tienen una puntuación mayor que cualquier jaula con la misma topología.
3. Tratar el caso cuando todas las caras de la jaula son triángulos y cuadriláteros, con la excepción del prisma pentagonal, que es más difícil.
4. Tratar cualquier jaula que tenga una cara con más de cuatro lados.
5. Resolver el único caso restante, cuando la jaula es un prisma pentagonal.

La parte 1 fue resuelta en 1994, y la parte 2 en 1995. A medida que el programa se desarrollaba, Hales modificó la definición de una jaula para simplificar el argumento (él la llamaba «estrella en descomposición»). La nueva definición no altera las dos jaulas ilustradas, y no tenía ningún efecto serio en aquellas partes de la demostración que ya se habían obtenido. En 1998, utilizando el nuevo concepto, se habían completado las cinco etapas. Un estudiante de Hales, Samuel Ferguson, resolvió la parte 5, el caso difícil de un prisma pentagonal.

El análisis incluía un uso intensivo de un ordenador en todas las etapas. El truco está en elegir, para cada red local, una noción de puntuación que haga el cálculo relativamente fácil. Desde el punto de vista geométrico, reemplazar la densidad por la puntuación es como poner una especie de tejado sobre la parte superior del suave paisaje cuya cima se está buscando. El tejado está hecho de

muchas piezas planas (véase Figura 24). Formas como esta son más fáciles de tratar que superficies lisas, porque los máximos deben ocurrir en las esquinas y Máximo del tejado estas pueden encontrarse resolviendo ecuaciones mucho más sencillas. Hay métodos eficientes para hacerlo, conocidos como programación lineal. Si el tejado ha sido astutamente construido de modo que su cima coincida con la cima de la superficie lisa, entonces este cálculo más sencillo localiza la cima de la superficie lisa.

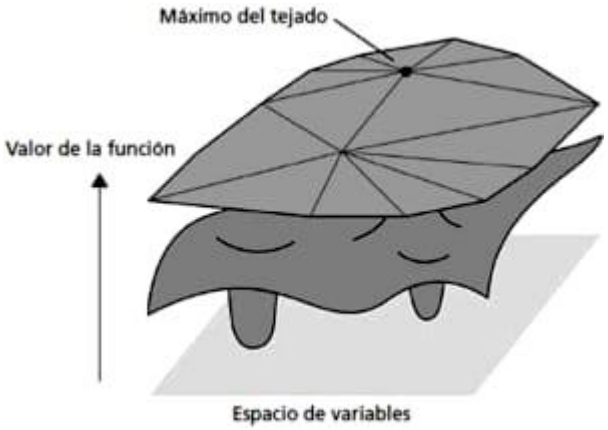


Figura 24. Encajar un tejado sobre la parte superior de una función.

Hay un precio que pagar por esta aproximación: hay que resolver unos cien mil problemas de programación lineal. Los cálculos son muy largos, pero están dentro de las capacidades de los ordenadores actuales. Cuando Hales y Ferguson prepararon su trabajo para su publicación ocupaban unas doscientas cincuenta páginas de matemáticas, más tres *gigabytes* de archivos de ordenador.

En 1999 Hales envió la demostración a *Annals of Mathematics*, y la revista escogió un panel de doce evaluadores expertos. Para 2003 el panel se declaró «un 99 por 100 seguro» de que la demostración era correcta. La incertidumbre restante concernía a los cálculos por ordenador; los miembros del panel habían repetido muchos de ellos, y en otros casos habían comprobado la forma en que la demostración estaba organizada y programada, pero fueron incapaces de verificar algunos puntos. Con alguna demora, la revista publicó el artículo. Hales reconocía que era probable que esta aproximación a la demostración nunca fuera certificada como correcta al cien por cien, y en 2003 anunció que estaba iniciando un proyecto para reformular la demostración de una manera que pudiera verificarse mediante un ordenador utilizando *software* estándar automatizado para comprobar demostraciones.

Esto puede sonar como salir de la sartén para caer en el fuego, pero en realidad tiene perfecto sentido. Las demostraciones que publican los matemáticos en las revistas pretenden convencer a los seres humanos. Como dije en el capítulo 1, este tipo de demostración es una especie de historia. Los ordenadores no son buenos contando historias, pero son excelentes en algo que nosotros somos inútiles: realizar cálculos largos y tediosos sin cometer errores. Los ordenadores son ideales para el concepto formal de demostración en los libros de texto: una serie de pasos lógicos, cada uno de los cuales se sigue de los anteriores.

Los científicos de la computación han explotado esta capacidad. Para comprobar una demostración, hacemos que un ordenador

verifique cada paso lógico. Debería ser fácil, pero las demostraciones en las revistas no están escritas así. Dejan fuera cualquier cosa rutinaria u obvia. Es fácil detectar frases consabidas: «Es fácil verificar que…», «Utilizando los métodos de Cheesburguer y Fries, modificados para tener en cuenta singularidades aisladas, vemos que…», «Un breve cálculo establece…». Los ordenadores no pueden (aún) manejar cosas de este tipo. Pero los seres humanos pueden reescribir las demostraciones llenando todos estas lagunas, y los ordenadores pueden entonces verificar cada paso.

La razón de que no estemos saltando sobre brasas es simple: el *software* que hace la verificación tiene que comprobarse solo *una vez*. Es *software* de propósito general, aplicable a todas las demostraciones escritas en el formato correcto. Todas las preocupaciones sobre las demostraciones por ordenador se centran en ese elemento de *software*. Verificado ese, puede ser utilizado para verificar todo lo demás. Incluso se puede hacer el proceso autoconsistente escribiendo el *software* de verificación de demostraciones en un lenguaje que pueda comprobarse mediante un *software* de verificación de demostraciones mucho más simple.

En años recientes se han verificado de esta manera las demostraciones de muchos teoremas matemáticos clave. A menudo las demostraciones tienen que presentarse en un estilo que es más apropiado para el tratamiento mediante ordenador. Uno de los éxitos actuales es una demostración verificada del teorema de la curva de Jordan: toda curva cerrada en el plano que no se cruza consigo misma divide al plano en dos regiones conexas diferentes.

Esto puede sonar obvio, pero los pioneros de la topología tuvieron problemas para encontrar una demostración rigurosa. Camille Jordan lo consiguió finalmente en 1887 con una demostración de más de ochenta páginas, pero más tarde fue criticado por hacer hipótesis injustificadas. En su lugar, el crédito se lo llevó Oswald Veblen, quien dio una demostración más detallada en 1905, donde decía que «la demostración [de Jordan]... es insatisfactoria para muchos matemáticos. Supone el teorema sin demostración en el caso especial importante de un simple polígono, y a partir de ese punto uno debe admitir al menos que no se dan todos los detalles del argumento». Más tarde los matemáticos aceptaron la crítica de Veblen sin poner reparos, pero recientemente Hales volvió sobre la demostración de Jordan y no encontró «nada objetable» en ella. El comentario de Veblen sobre un polígono es extraño: el teorema es sencillo para un polígono, y en cualquier caso la demostración de Jordan no se basa en esta versión[46]. Las demostraciones narrativas

[46] Supongamos que la región es un polígono, como en la Figura 55. Dado cualquier punto que no está en el polígono, existe una línea recta desde dicho punto que sale fuera de un gran círculo que contiene al polígono y no pasa por ningún vértice del polígono. (Existe un número finito de vértices pero infinitas líneas rectas para escoger). Esta línea corta al polígono un número finito de veces, y este número es par o impar. Definimos el interior como lo que consiste en todos los puntos para los cuales el número es impar, y el exterior como lo que consiste en todos los puntos para los que es par. Es entonces sencillo demostrar que cada una de estas regiones es conexa y el polígono las separa.

tienen sus propios peligros. Siempre vale la pena comprobar si la versión popular de la historia es la misma que la original.

Como calentamiento para la conjetura de Kepler, Hales dio en 2007 una demostración formal verificada por ordenador del teorema de la curva de Jordan, que utilizaba sesenta mil líneas de programa informático. Poco después, un equipo de matemáticos dio otra demostración formal utilizando un *software* diferente. La verificación por ordenador no está por completo a salvo de fallos, pero tampoco lo están las demostraciones tradicionales. De hecho, es probable que muchos artículos de investigación en matemáticas contengan un error técnico en alguna parte. Estos errores se manifiestan ocasionalmente, y la mayoría de ellos resultan ser inocuos. Los errores graves son detectados por lo general porque introducen inconsistencias tales que hay algo que visiblemente no tiene sentido. Esta es otra desventaja de la aproximación narrativa: el precio que pagamos por hacer que una demostración sea

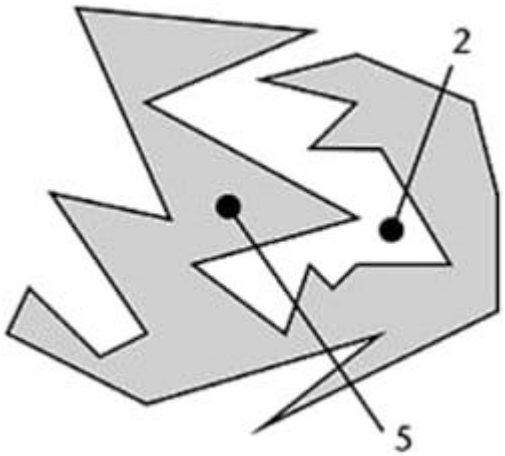


Figura 55. Demostrando el teorema de la curva de Jordan para un polígono. Un número impar de intersecciones ocurre para puntos en la región sombreada (interior), y un número par de intersecciones ocurre para puntos en la región blanca (exterior).

comprensible por los seres humanos es que una historia entretenida puede a veces ser muy convincente incluso si es falsa.

Hales llama a su aproximación *Proyecto FlysPecK* —las F, P y K significan «prueba formal de Kepler»—. Inicialmente él estimó que le llevaría unos veinte años completar la tarea[47]. Cuando el proyecto llevaba nueve años ya se habían hecho avances considerables. Quizá acabe pronto.

[47] http://code.google.com/p/flyspeck/.

Capítulo 6

Nuevas soluciones para lo viejo

La conjetura de Mordell

Ahora volvemos a entrar en los dominios de la teoría de números, teniendo como objetivo el último teorema de Fermat. Para preparar el terreno empezaré con un problema menos familiar pero presumiblemente aún más importante. En 2002 Andrew Granville y Thomas Tucker lo presentaban así[48]:

> *En [1922] Mordell escribió uno de los más grandes artículos en la historia de las matemáticas... En el comienzo del artículo Mordell planteaba cinco preguntas que fueron instrumentales para motivar buena parte de la investigación importante en aritmética diofántica en el siglo XX. La más importante y difícil de estas preguntas fue respondida por Faltings en 1983 inventando algunas de las ideas más profundas y poderosas en la historia de las matemáticas.*

Mordell es el teórico de números británico Louis Mordell, quien había nacido en Estados Unidos dentro de una familia judía de origen lituano, y Faltings es el matemático alemán Gerd Faltings. La pregunta mencionada llegó a conocerse como conjetura de Mordell, y la cita da su estatus actual: demostrada, brillantemente, por Faltings.

---

[48] Andrew Granville y Thomas Tucker, «It's as easy as abc», *Notices of the American Mathematical Society*, n.º 49 (2002), pp. 1224-1231.

La conjetura de Mordell pertenece a un área principal de la teoría de números: las ecuaciones diofánticas. Se llaman así por Diofanto de Alejandría, quien escribió un libro famoso, *Arithmetica*, alrededor de 250 d. C. Se cree que originalmente la *Arithmetica* contenía trece libros, pero solo seis han sobrevivido, todos ellos en copias posteriores. Este no era un texto de aritmética en el sentido de sumas y multiplicaciones. Fue el primer texto de álgebra y recogía casi todo lo que los griegos sabían sobre la forma de resolver ecuaciones. Incluso contenía una forma rudimentaria de notación algebraica, que se cree que utilizaba una variante σ de la letra griega sigma para la incógnita (nuestra x), Δ^{Y} para su cuadrado (nuestra x^2) y K^{Y} para su cubo (nuestra x^3). La suma se denotaba colocando unos símbolos a continuación de otros, la resta tenía su propio símbolo especial, el recíproco de la incógnita (nuestro $1/x$) era σ^{x}, y había otras notaciones de este tipo. Los símbolos han sido reconstruidos a partir de copias y traducciones posteriores, y quizá no sean completamente exactos.

En el espíritu de las matemáticas griegas clásicas, se exigía que las soluciones de las ecuaciones que buscaba la *Arithmetica* fueran números racionales, es decir, fracciones tales como 22/7 formadas con números enteros. A menudo se exigía que las propias soluciones *fueran* números enteros. Todos los números implicados eran positivos; los números negativos fueron introducidos varios siglos más tarde en China y la India. Ahora llamamos a estos problemas ecuaciones diofánticas. El libro incluye algunos resultados notablemente profundos. En particular, parece que

Diofanto es consciente de que todo número entero puede expresarse como suma de cuatro cuadrados perfectos (incluido el cero). Lagrange dio la primera demostración en 1700. El resultado que aquí nos interesa es una fórmula para todas las tripletas pitagóricas, en donde dos cuadrados perfectos se suman para dar otro cuadrado perfecto. El nombre procede del teorema de Pitágoras: esta relación es válida para los lados de un triángulo rectángulo. El ejemplo más conocido es el famoso triángulo 3 - 4- 5: $3^2 + 4^2 = 5^2$. Otro es $5^2 + 12^2 = 13^2$. Hay infinitas tripletas pitagóricas, y hay una receta para encontrarlas en dos lemas (proposiciones auxiliares) que preceden a las proposiciones 29 y 30 del libro X de los *Elementos* de Euclides.

El procedimiento de Euclides da infinitas tripletas pitagóricas. Mordell conocía otras ecuaciones diofánticas para las que existe una fórmula que da infinitas soluciones. También conocía otro tipo de ecuación diofántica con infinitas soluciones, no prescritas por una fórmula. Estas se denominan curvas elípticas —un nombre bastante estúpido puesto que prácticamente no tienen nada que ver con elipses— y la infinidad de soluciones aparece porque dos soluciones cualesquiera pueden combinarse para dar otra solución. El propio Mordell demostró una de las propiedades básicas de estas ecuaciones: solo se necesita un número finito de soluciones para generar todas las demás mediante este proceso.

Aparte de estos dos tipos de ecuaciones, cualquier otra ecuación diofántica en la que Mordell podía pensar caía en una de dos categorías. O bien se sabía que tenía solo un número finito de

soluciones, incluida ninguna, o bien nadie sabía si el número de soluciones era finito o infinito. Esto por sí solo no era noticia, pero Mordell pensó que podía detectar una pauta que nadie más había advertido. No era una pauta de teoría de números; procedía de la topología. Lo que importaba era cuántos agujeros tenía la ecuación. Y para dar sentido a esto, había que considerar sus soluciones en números complejos, no números racionales o enteros. Lo que de algún modo parecía contrario al espíritu mismo de las ecuaciones diofánticas.

Vale la pena exponer aquí algunos detalles que serán de ayuda más tarde. No se asuste por el álgebra; está aquí básicamente para darme algo concreto a lo que referirme. Concentrémonos en la historia que hay detrás del álgebra.

Las tripletas pitagóricas son soluciones, en números enteros, de la ecuación pitagórica

$$x^2 + y^2 = z^2$$

Dividiendo por z^2 da

$$(x/z)^2 + (y/z)^2 = 1$$

Según el capítulo 3, esto nos dice que el par de números racionales $(x/z, y/z)$ se encuentra sobre el círculo unidad en el plano. Ahora bien, la ecuación pitagórica tenía su origen en la geometría, y su interpretación es que el triángulo asociado tiene un ángulo recto. La

fórmula que acabo de obtener proporciona una interpretación geométrica algo diferente, no solo de una tripleta pitagórica sino de todas ellas. Las soluciones de la ecuación pitagórica corresponden directa y naturalmente a todos los puntos racionales sobre el círculo unidad. Aquí se dice que un punto es racional siempre que lo sean sus dos coordenadas.

Pueden deducirse muchos hechos interesantes a partir de esta relación. Con un poco de trigonometría, o por álgebra directa, puede descubrirse que para cualquier número *t* el punto

$$\left(\frac{2t}{t^2+1}, \frac{t^2-1}{t^2+1}\right)$$

se encuentra sobre el círculo unidad. Además, si *t* es racional, también lo es este punto. Todos los puntos racionales surgen de esta manera, de modo que tenemos una fórmula completa para todas las soluciones de la ecuación pitagórica. Es equivalente a la fórmula de Euclides, que es la misma que la de Diofanto. Como ejemplo, si $t = 22/7$ entonces la fórmula da

$$\left(\frac{308}{533}, \frac{435}{533}\right)$$

y es fácil comprobar que $308^2 + 435^2 = 533^2$. Para nosotros, la fórmula exacta no es terriblemente importante; lo que importa es que existe una.

Esta no es la única ecuación diofántica para la que una fórmula da todas las soluciones, pero son relativamente raras. Otras incluyen las denominadas ecuaciones de Pell, tales como $x^2 = 2y^2 + 1$. Esta tiene infinitas soluciones, tales como $3^2 = 2\times2^2 + 1$, $17^2 = 2\times12^2 + 1$, y existe una fórmula general. Sin embargo, las tripletas pitagóricas tienen más estructura que eso, derivada también de la geometría. Supongamos que tenemos dos tripletas pitagóricas. Entonces hay dos soluciones correspondientes de la ecuación pitagórica —puntos racionales en el círculo—. La geometría ofrece una forma natural de «sumar» dichos puntos. Partimos del punto (1,0) en el que el círculo corta al eje horizontal, y encontramos los ángulos entre este punto y las dos soluciones. Sumamos los dos ángulos (véase Figura 25), y constatamos qué punto resulta. Ciertamente se encuentra sobre el círculo. Un breve cálculo muestra que es racional. De modo que a partir de dos soluciones cualesquiera podemos obtener una tercera. Los matemáticos ya habían advertido muchos hechos como este. La mayoría tienen sentido inmediato si se piensa en los puntos racionales sobre un círculo.

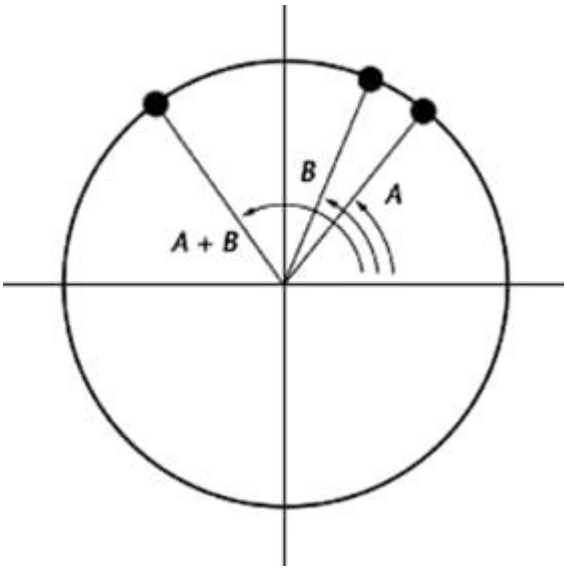


Figura 25. Combinando dos soluciones racionales A y B de la ecuación pitagórica para obtener una tercera, A + B.

El «breve cálculo» que he mencionado hace uso de la trigonometría. Las funciones trigonométricas clásicas tales como el seno y el coseno están íntimamente relacionadas con la geometría de un círculo. El cálculo aludido utiliza fórmulas estándar bastante elegantes para el seno y el coseno de la suma de dos ángulos en términos de los senos y cosenos de los propios ángulos. Hay muchas maneras de construir senos y cosenos, y una bastante clara procede del cálculo integral. Si se integra la función algebraica $1/\sqrt{1x^2}$, el resultado puede expresarse en términos de la función seno. En realidad, lo que necesitamos es la función inversa del seno: el ángulo cuyo seno es el número en el que estamos pensando[49].

[49] Para desarrollar este críptico comentario: la fórmula es

$$\int \frac{dx}{\sqrt{1-x^2}} = \operatorname{arcsen} x$$

La integral aparece cuando tratamos de derivar una fórmula para la longitud de un arco de círculo utilizando el cálculo infinitesimal, y la geometría del círculo tiene una implicación sencilla pero muy importante para el resultado. La circunferencia del círculo unidad es 2π, de modo que yendo alrededor del círculo una distancia 2π volvemos exactamente al mismo punto. Lo mismo sucede con cualquier múltiplo entero de 2π: por el convenio matemático estándar, los enteros positivos corresponden al sentido contrario a las agujas del reloj, y los negativos al sentido de las agujas del reloj. De ello se sigue que el seno y el coseno de un número permanecen invariables si a dicho número se le suma un múltiplo entero de 2π. Decimos que la función es periódica, con período 2π.

Los analistas de los siglos XVIII y XIX descubrieron una amplia generalización de esta integral, junto con muchas nuevas funciones interesantes análogas a las familiares funciones trigonométricas. Estas nuevas funciones eran intrigantes: eran periódicas, como el seno y el coseno, pero de un modo más sofisticado. En lugar de tener un período, como 2π (y sus múltiplos enteros), tenían dos períodos independientes. Si tratamos de hacer esto con funciones reales, todo lo que obtenemos son constantes, pero en el caso de funciones complejas las posibilidades son mucho más ricas.

El área fue iniciada por el matemático italiano Giulio di Fagnano y el prolífico Euler. Fagnano estaba tratando de encontrar la longitud del arco de una elipse utilizando el cálculo infinitesimal, pero no pudo encontrar una fórmula explícita —lo que ya no sorprende

---

donde arcsen (a menudo escrito sen^{-1}) es la función inversa del seno. Es decir, si y = sen x, entonces x = arcsen y.

puesto que ahora sabemos que no existe—. Sin embargo, advirtió una relación entre las longitudes de varios arcos especiales y la publicó en 1750. Euler advirtió la misma relación en el mismo contexto y la presentó como una relación formal entre integrales. Son similares a la asociada con la función seno, pero la expresión cuadrática $1 - x^2$ bajo la raíz cuadrada está reemplazada por un polinomio cúbico o cuártico, por ejemplo el polinomio cuártico $(1 - x^2)(1 - 4x^2)$.

En 1811 Adrien-Marie Legendre publicó el primer libro de un voluminoso tratado en tres volúmenes sobre estas integrales, que se conocen como integrales elípticas debido a su conexión con la longitud de arco de un segmento de elipse. Sin embargo, él pasó por alto la propiedad más importante de estas integrales: la existencia de nuevas funciones, análogas a las funciones seno y coseno, cuyas funciones *inversas* expresan el valor de la integral de una manera sencilla[50]. Gauss, Niels Henrik Abel y Carl Jacobi rápidamente detectaron el descuido. Gauss, como era habitual en él, se guardó el descubrimiento para sí mismo. Abel envió un artículo a la Academia Francesa en 1826, pero Cauchy, el presidente, extravió el manuscrito y no fue publicado hasta 1841, doce años después de la temprana y trágica muerte de Abel por tuberculosis. Sin embargo, otro artículo de Abel sobre el mismo tema se publicó en 1827.

[50] Por ejemplo, sea k un número complejo cualquiera y consideremos la integral

$$\int \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}}$$

Esta es la función inversa de una función elíptica denotada por sn. Existe una función semejante para cada valor de k. La situación es como la de la nota 49, pero más elaborada.

Jacobi hizo de estas nuevas «funciones elípticas» la base en un enorme volumen, publicado en 1829, que impulsó el análisis complejo en una trayectoria completamente nueva.

Lo que emergió fue un bello paquete de propiedades interrelacionadas, análogas a las de las funciones trigonométricas. La relación advertida por Fagnano y Euler podía reinterpretarse como una simple lista de fórmulas que relacionan funciones elípticas de la suma de dos números con funciones elípticas de los propios números. La propiedad más maravillosa de las funciones elípticas supera a las funciones trigonométricas de una forma espectacular. Las funciones elípticas no solo son periódicas: son doblemente periódicas. Una línea es unidimensional, de modo que solo pueden repetirse pautas en una dirección, a lo largo de la línea. El plano complejo es bidimensional, de modo que las pautas pueden repetirse como en el papel de pared: a lo largo del rollo de papel y también a lo largo de la pared en franjas de papel adyacentes. Asociados con cada función elíptica hay dos números complejos independientes, sus períodos, y sumar cualquiera de ellos a la variable no cambia el valor de la función.

Repitiendo este proceso concluimos que el valor de la función no cambia si sumamos a la variable cualquier combinación entera de los dos períodos. Estas combinaciones tienen una interpretación geométrica: determinan una red en el plano complejo. La red especifica una teselación del plano por paralelogramos, y cualquier cosa que suceda en un paralelogramo tiene una copia en todos los demás (véase Figura 26). Si consideramos solo un paralelogramo, la

forma en que se une a copias adyacentes significa que tenemos que identificar lados opuestos, de la misma forma que un toro se define identificando lados opuestos de un cuadrado (véase Figura 12). Un paralelogramo con lados opuestos identificados es también un toro topológico. Así como el seno y el coseno están relacionados con el círculo, las funciones elípticas están relacionadas con un toro.

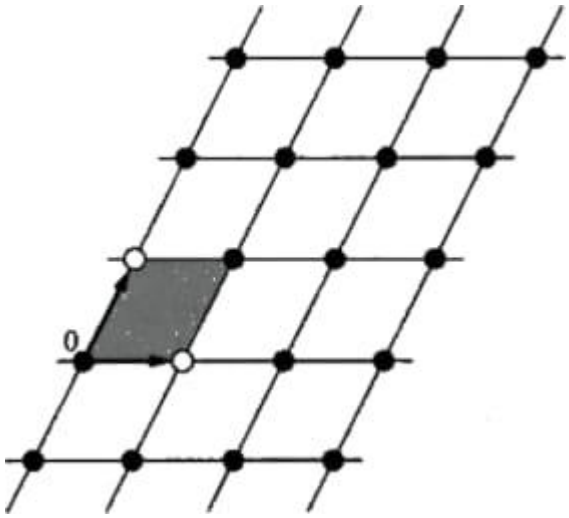


Figura 26. Red en el plano complejo. Las flechas apuntan a los dos períodos, mostrados como puntos blancos. El valor de la función en el paralelogramo sombreado la determina en cualquier otro paralelogramo.

También hay un vínculo con la teoría de números. Dije que la función inversa del seno se obtiene integrando una fórmula que incluye a la raíz cuadrada de un polinomio cuadrático. Las funciones elípticas son similares, pero el polinomio cuadrático está reemplazado por un polinomio cúbico o cuártico. El caso cuártico ha sido mencionado antes de pasada, porque históricamente llegó

primero, pero ahora vamos a centrarnos en el caso cúbico. Si denotamos la raíz cuadrada por y, y el polinomio por $ax^3 + bx^2 + cx + d$ donde a, b, c, d son coeficientes numéricos, entonces x e y satisfacen la ecuación

$$y^2 = ax^3 + bx^2 + cx + d$$

Esta ecuación puede considerarse en varios contextos diferentes, dependiendo de qué restricciones se impongan sobre las variables y los coeficientes. Si son reales, la ecuación define una curva en el plano. Si son complejos, los geómetras algebraicos siguen llamando al conjunto de soluciones una curva, por analogía. Pero ahora es una curva en el espacio de pares de números complejos, que es tetradimensional en coordenadas reales. Y desde este punto de vista de números reales, la curva es en realidad una superficie.

La Figura 27 muestra las curvas elípticas reales $y^2 = 4x^3 - 3x + 2$ e $y^2 = 4x^3 - 3x$, que son típicas. Puesto que y aparece elevado al cuadrado, la curva es simétrica respecto al eje horizontal. Dependiendo de los coeficientes, es o bien una única curva sinuosa o tiene un componente oval separado. Sobre los números complejos, la curva es siempre una pieza conexa.

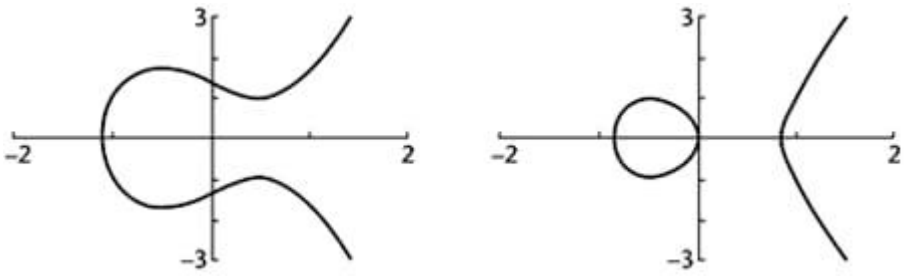


Figura 27. Curvas elípticas reales típicas. Izquierda: $y^2 = 4x^3 - 3x + 2$. Derecha: $y^2 = 4x^3 - 3x$.

La teoría de números entra en juego cuando exigimos que variables y coeficientes sean racionales. Ahora estamos examinando una ecuación diofántica. Debido al vínculo con las funciones elípticas, recibe el nombre bastante equívoco de curva elíptica, pese a que no se parece en nada a una elipse. Es como llamar círculo a una curva triangular debido al vínculo con la trigonometría. Por desgracia, el nombre está ahora grabado en tablas de piedra, de modo que tenemos que vivir con ello.

Puesto que las funciones elípticas tienen una teoría rica y profunda, los teóricos de números han descubierto muchas bellas propiedades de las curvas elípticas. Una es estrechamente análoga a la forma en que podemos combinar dos soluciones de la ecuación pitagórica sumando los ángulos asociados. Dos puntos en una curva elíptica pueden combinarse trazando una línea recta que pasa por ellos y viendo donde corta a la curva por tercera vez (véase Figura 28). (Siempre existe este tercer punto, porque la ecuación es una cúbica. Sin embargo, podría estar «en el infinito», o podría coincidir con uno de los dos primeros puntos si la recta es tangente a la curva). Si los dos puntos son P y Q, denotamos el tercero por $P*Q$.

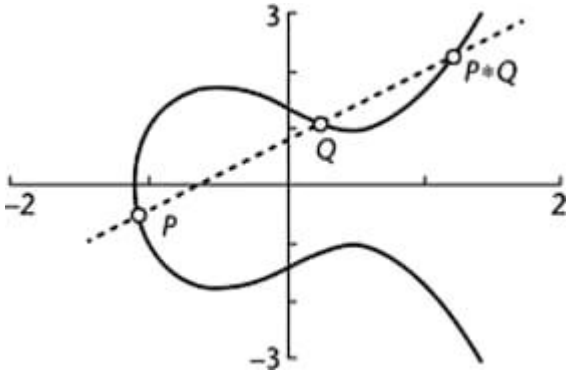


*Figura 28. Combinación de los puntos P, Q para obtener el punto P*Q.*

Un cálculo muestra que si *P* y *Q* son puntos racionales, entonces también lo es *P*Q*. La operación * da al conjunto de números racionales una estructura algebraica, pero resulta útil considerar una operación relacionada. Escojamos cualquier punto racional *O* en la curva y definamos

$$P + Q = (P * Q) * O$$

Esta nueva operación obedece a algunas leyes básicas del álgebra ordinaria, con *O* comportándose como cero, y convierte el conjunto de todos los puntos racionales en lo que los algebristas llaman un grupo (véase capítulo 10). El punto esencial es que, como las tripletas pitagóricas, se pueden «sumar» dos soluciones cualesquiera para obtener una tercera. Que se dé esta «ley de grupo» en los puntos racionales es sorprendente, y en particular significa que una vez que hemos encontrado dos soluciones racionales de la ecuación diofántica, automáticamente obtenemos muchas más.

Alrededor de 1908 Poincaré se preguntó si existe un número finito de soluciones a partir de las cuales pueden obtenerse *todas* las demás soluciones aplicando la operación de grupo una y otra vez. Este resultado es importante porque implica que todas las soluciones racionales pueden ser caracterizadas escribiendo una lista finita. En su espectacular artículo de 1922 Mordell demostró que la respuesta a la pregunta de Poincaré es «sí». Ahora las curvas elípticas adquieren una importancia fundamental en la teoría de números, porque no era habitual tener ese tipo de control sobre cualquier ecuación diofántica.

Tanto la ecuación pitagórica como las curvas elípticas tienen entonces infinitas soluciones racionales. Por el contrario, muchas ecuaciones diofánticas tienen solo un número finito de soluciones, y a veces ninguna. Voy a hacer una pequeña digresión para discutir toda una familia de tales ecuaciones y la reciente y notable demostración de que las soluciones obvias son las únicas que existen.

Los pitagóricos estaban interesados en su ecuación porque creían que el universo está basado en números. En apoyo de esta filosofía, ellos descubrieron que razones numéricas simples rigen la armonía musical. Lo observaron de manera experimental utilizando una cuerda tensa. Una cuerda de la misma tensión que tiene la mitad de su longitud emite una nota una octava más alta. Esta es la combinación más armoniosa de dos notas: tan armoniosa que suena un poco insulso. En la música occidental las siguientes armonías más importantes son la cuarta, donde una cuerda tiene

3/4 de la longitud de la otra, y la quinta, donde una cuerda tiene 2/3 de la longitud de la otra[51].

Partiendo de 1 y multiplicándolo repetidamente por 2 o 3 se obtienen los números 2, 3, 4, 6, 8, 9, 12 y así sucesivamente: números de la forma 2^a3^b. Debido a la conexión musical estos llegaron a conocerse como números armónicos. En el siglo XIII un escritor judío que vivía en Francia escribió *Sha'ar ha-Shamayim* («La puerta del cielo»), una enciclopedia basada en fuentes árabes y griegas. La dividió en tres partes: física, astronomía y metafísica. Su nombre era Gerson ben Solomon Catalan. En 1343 el obispo de Meaux convenció al hijo de Gerson (bueno, los historiadores piensan que probablemente era su hijo) Levi ben Gerson para escribir un libro matemático, *The Harmony of Numbers*. Incluía un problema planteado por el compositor y teórico de la música Philippe de Vitry: ¿cuándo pueden dos números armónicos diferir en 1? Es fácil encontrar pares semejantes: De Vitry conocía cuatro, a saber, (1,2), (2,3), (3,4) y (8,9). Ben Gerson demostró que estas son las únicas soluciones posibles.

Entre los pares de números armónicos de De Vitry, el más interesante es (8, 9). El primero es un cubo, 2^3; el segundo es un cuadrado, 3^2. Los matemáticos empezaron a preguntarse si otros cuadrados y cubos podrían diferir en 1, y Euler demostró que no podían hacerlo, aparte del caso trivial (0, 1), y también (–1, 0) si se admiten números negativos. En 1844 el segundo Catalan en la historia decidió llevar a la prensa una afirmación más radical, que

[51] Véase Ian Stewart, *Seventeen Equations that Changad the World*, Profile, 2012, capítulo 8.

muchos matemáticos debían haber pensado pero no se habían molestado en hacer explícita. Era el matemático belga Eugène Charles Catalan, y en 1844 escribió a una de las revistas matemáticas destacadas de la época, el *Journal für die Reine und Argewandte Mathematik* («Revista de matemáticas puras y aplicadas»):

> *Le ruego, señor, haga el favor de anunciar en su revista el siguiente teorema que creo cierto aunque aún no he conseguido demostrarlo por completo; quizá otros tengan más éxito. Dos números naturales consecutivos, aparte de 8 y 9, no pueden ser potencias consecutivas; dicho de otra manera, la ecuación* $x^m - y^n = 1$ *en la que las incógnitas son enteros positivos solamente admite una única solución.*

Esta proposición llegó a conocerse como la conjetura de Catalan. Los exponentes m y n son enteros mayores que 1.

Pese a progresos parciales, la conjetura de Catalan se negaba con obstinación a rendirse, hasta que fue espectacularmente demostrada en 2002 por Preda Mihăilescu. Nacido en Rumanía en 1955, se había establecido en Suiza en 1973 y acababa de completar su doctorado. El título de su tesis era «Ciclotomía de anillos y comprobación de primalidad», y aplicaba la teoría de números a la comprobación de la primalidad (véase capítulo 2). Este problema no tenía ninguna relación especial con la conjetura de Catalan, pero Mihăilescu llegó a advertir que casi con certeza sus métodos sí la tenían. Se derivaban de las ideas que mencioné en el

capítulo 3: la construcción de Gauss del 17-gono regular y las ecuaciones algebraicas asociadas, cuyas soluciones se denominan números ciclotómicos. La demostración era muy técnica y causó impresión en la comunidad matemática. Nos dice que cualesquiera que sean los valores que escojamos para las dos potencias, el número de soluciones es finito; y salvo las soluciones obvias que utilizan 0 y 1, la única interesante es $3^2 - 2^3 = 1$.

Los ejemplos anteriores muestran que algunas ecuaciones diofánticas tienen infinitas soluciones, y otras no. ¡Perfecto!: estas alternativas lo cubren todo. No obstante, si empezamos a preguntar qué ecuaciones son de cada tipo la cosa se hace más interesante. Mordell, un experto en ecuaciones diofánticas, estaba escribiendo un libro de texto seminal. En su época el área se parecía a la biología primitiva: mucho coleccionismo de mariposas y muy poca clasificación sistemática. Un Pitagórico Pintado aquí, una Gran Elíptica Azul allá, y en los arbustos orugas del Pelliano Moteado. El campo estaba prácticamente tal como lo había dejado Diofanto: una lista desestructurada de trucos separados, uno por cada tipo de ecuación. Este es un pobre material para un libro de texto, y pedía a gritos una organización. Eso es justamente lo que se propuso hacer Mordell.

En algún momento él debió advertir que todas las ecuaciones de las que se sabe que tienen infinitas soluciones racionales —tales como la ecuación pitagórica y las curvas elípticas— tenían una característica común. Él se centró en una clase de ecuaciones, aquellas que (después de ser convertidas en ecuaciones en números

racionales, como yo hice en el caso de Pitágoras) incluyen solo dos variables. Hay dos casos en los que sabemos cómo encontrar infinitas soluciones. Uno está ejemplificado por la ecuación pitagórica en la forma equivalente $x^2 + y^2 = 1$. Aquí hay una fórmula para las soluciones. Introduzcamos cualquier número racional en la fórmula y obtenemos una solución racional, y así aparecen todas las soluciones. La otra está ejemplificada por las curvas elípticas: existe un *proceso* que genera nuevas soluciones a partir de las viejas, y hay una garantía de que si se empieza con un conjunto finito apropiado de soluciones, este proceso las genera todas.

La conjetura de Mordell afirma que cuando quiera que haya infinitas soluciones racionales, debe aplicarse una de estas dos propiedades. O bien hay una fórmula general, o bien hay un proceso que genera todas las soluciones a partir de un apropiado conjunto finito de ellas. En todos los demás casos, el número de soluciones racionales es finito, por ejemplo las ecuaciones $x^m - y^n = 1$ que intervienen en la conjetura de Catalan. En cierto sentido, las soluciones son entonces meras coincidencias, sin ninguna estructura subyacente.

Mordell llegó a esta observación de una manera algo diferente. Él advirtió que toda ecuación con infinitas soluciones racionales tiene una sorprendente propiedad topológica. Tiene género 0 o 1. Recordemos del capítulo 4 que el género es un concepto de la topología de superficies y cuenta cuántos agujeros tiene la superficie. Una esfera tiene género 0, un toro género 1, un toro con dos agujeros tiene género 2, y así sucesivamente. ¿Cómo entran las

superficies en un problema de teoría de números? A partir de la geometría de coordenadas. Vimos que la ecuación pitagórica, interpretada en términos de números racionales y ampliada para admitir números reales como soluciones, determina un círculo. Mordell fue un paso más lejos y admitió números complejos como soluciones. Cualquier ecuación con dos variables complejas determina lo que los geómetras algebraicos llaman una curva compleja. Sin embargo, desde el punto de vista de los números reales y del sistema visual humano, todo número complejo es bidimensional: tiene dos componentes reales, su parte real y su parte imaginaria. De modo que a ojos complejos una «curva» es una superficie para usted y para mí. Siendo una superficie, tiene un género: ahí lo tiene.

En el caso de las curvas de las que se sabía que solo tienen soluciones finitas, su género era al menos 2. Ecuaciones importantes cuyo estatus era desconocido también tenían género al menos 2. En un salto aventurado y valiente basado en lo que entonces parecía evidencia muy endeble, Mordell conjeturó que cualquier ecuación diofántica con género 2 o mayor solamente tiene un número finito de soluciones racionales. De golpe, las mariposas diofánticas estaban limpiamente dispuestas en familias emparentadas; de forma adecuada, por género.

Había solo una minúscula pega en la conjetura de Mordell. Relacionaba dos cosas muy diferentes: soluciones racionales y topología. En esa época, cualquier vínculo plausible era extremadamente tenue. Si existía una conexión, nadie sabía cómo

encontrarla. De modo que la conjetura era una especulación aventurada e injustificada, pero el beneficio potencial era enorme.

En 1983 Faltings publicó una espectacular demostración de que la especulación aventurada de Mordell estaba en lo cierto. Su demostración utilizaba métodos profundos de geometría algebraica. Una demostración muy diferente, basada en aproximar números reales por números racionales, fue pronto encontrada por Paul Votja, y Enrico Bombieri publicó una demostración simplificada siguiendo las mismas líneas en 1990. Hay una aplicación del teorema de Faltings al último teorema de Fermat, un problema que trataremos ampliamente en el capítulo 7. Esta afirma que para cualquier entero n mayor o igual que 3, la ecuación $x^n + y^n = 1$ tiene solo un número finito de soluciones enteras. El género de la curva asociada es $(n - 1)(n - 2)/2$, y esto es al menos 3 si n es 4 o mayor. El teorema de Faltings implica inmediatamente que para cualquier $n \geq 4$, la ecuación de Fermat tiene como máximo un número finito de soluciones racionales. Fermat afirmaba que no tenía ninguna excepto cuando x o y es cero, de modo que esto fue un gran avance. En el próximo capítulo abordaremos la historia del último teorema de Fermat y veremos cómo la afirmación de Fermat fue completamente vindicada.

Capítulo 7

Márgenes estrechos

El último teorema de Fermat

Encontramos por primera vez a Fermat en el capítulo 2, donde su elegante teorema sobre potencias de números proporcionaba un método para comprobar si un número es primo. Este capítulo trata una afirmación mucho más difícil: el último teorema de Fermat. Suena muy misterioso. «Teorema» parece claro, pero ¿quién era Fermat y por qué este era su *último* teorema? ¿Es el nombre de una astuta estrategia de *marketing*? No lo es: el nombre quedó asociado al problema en el siglo XVIII, cuando solo unos pocos matemáticos destacados habían oído hablar del mismo o se preocupaban por ello. Pero el último teorema de Fermat es realmente misterioso.

Pierre Fermat nació en Francia en 1601, según algunas fuentes, y en 1607-1608, según otras. La discrepancia quizá sea debida a una confusión con un primo del mismo nombre. Su padre era un rico mercader de cuero y ocupaba un alto cargo en el gobierno local, y su madre procedía de una familia de abogados del Parlamento. Él fue a la Universidad de Toulouse, se trasladó a Burdeos a finales de la década de 1620 y allí mostró signos prometedores de talento matemático. Hablaba varios idiomas, y preparó una recuperación de una obra perdida de las matemáticas griegas debida a Apolonio. Compartió sus muchos descubrimientos con destacados matemáticos de la época.

En 1631, tras graduarse en derecho en la Universidad de Orléans, fue nombrado consejero del Alto Tribunal de la Judicatura en Toulouse. Esto le autorizaba a cambiar su apellido por «de Fermat», y siguió siendo consejero por el resto de su vida. Su pasión, sin embargo, eran las matemáticas. Publicó poco, y prefería escribir cartas esbozando sus descubrimientos, por lo habitual sin demostración. Su trabajo era debidamente reconocido por los profesionales, con muchos de los cuales tenía estrechas relaciones, aunque conservaba su estatus *amateur*. Pero Fermat tenía tanto talento que era, de hecho, un profesional; solo que no ocupaba una posición oficial en las matemáticas.

Algunas de sus demostraciones han sobrevivido en cartas y artículos, y es evidente que Fermat sabía en qué consistía una genuina demostración. Después de su muerte, muchos de sus más profundos teoremas seguían sin estar demostrados, y los profesionales empezaron a trabajar sobre ellos. En menos de unas décadas se habían demostrado todos los enunciados de Fermat salvo uno, de modo que fue algo natural que este llegara a conocerse como su último teorema. A diferencia de los demás, se resistió a sucumbir, y pronto se hizo tristemente famoso por el contraste entre la simplicidad de su enunciado y la evidente dificultad de encontrar una demostración.

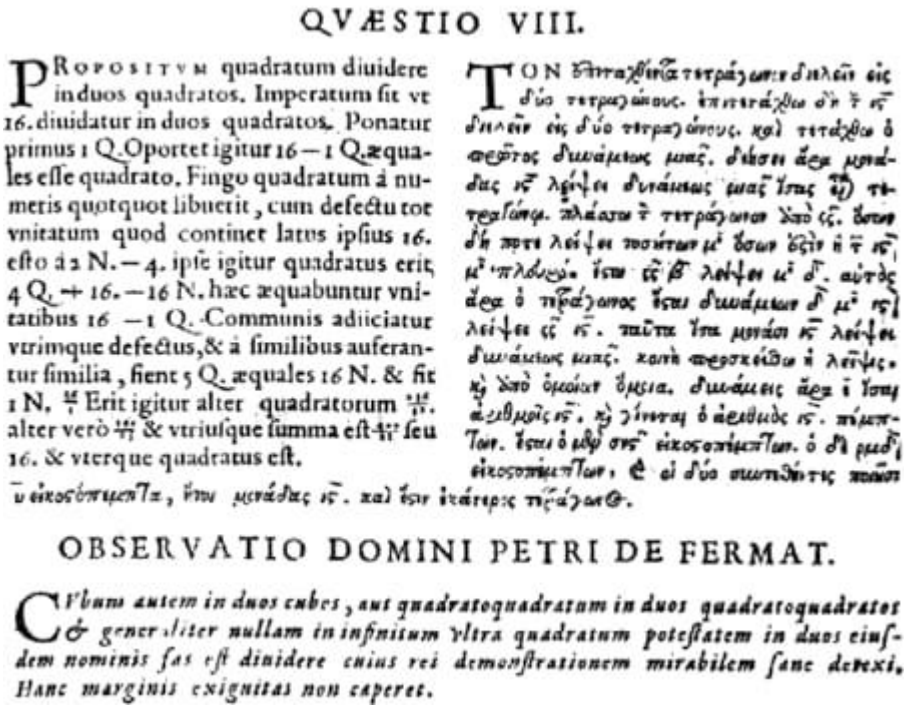
Parece que Fermat conjeturó su teorema alrededor de 1630. No se conoce la fecha exacta, pero fue cuando Fermat empezó a leer una edición recientemente publicada de la *Arithmetica* de Diofanto. Y es de ahí de donde sacó la idea. El último teorema vio la imprenta por

primera vez en 1670, cinco años después de la muerte de Fermat, cuando su hijo Samuel publicó una edición de la *Arithmetica*. Esta edición tenía una característica novedosa. Incorporaba las notas que Pierre había escrito en los márgenes de su copia personal de la traducción latina de Claude Gaspard Bacher de Méziriac de 1621. El último teorema está enunciado como una nota añadida a la cuestión VII del libro II de Diofanto (véase Figura 29).

El problema allí resuelto consiste en escribir un cuadrado perfecto como suma de dos cuadrados perfectos. En el capítulo 6 vimos que existe un número infinito de estas tripletas pitagóricas. Diofanto plantea una pregunta relacionada pero más difícil: cómo encontrar los dos lados más pequeños de un triángulo, dado el más largo. Un cuadrado específico debe ser «dividido» en dos cuadrados, es decir, debe ser expresado como su suma. Él muestra cómo resolver este problema cuando el lado mayor del triángulo es 4, y obtiene la respuesta

$$4^2 = (16/5)^2 + (12/5)^2$$

en números racionales. Multiplicando ambos miembros por 25 obtenemos $20^2 = 16^2 + 12^2$, y dividiendo por 16 obtenemos la familiar $3^2 + 4^2 = 5^2$. Normalmente Diofanto ilustraba los métodos generales con ejemplos concretos, una tradición que se remonta a la antigua Babilonia, y no daba demostraciones.



QVÆSTIO VIII.

PROPOSITVM quadratum diuidere in duos quadratos. Imperatum sit vt 16. diuidatur in duos quadratos. Ponatur primus 1 Q. Oportet igitur 16 — 1 Q. æquales esse quadrato. Fingo quadratum à numeris quotquot libuerit, cum defectu tot vnitatum quod continet latus ipsius 16. esto à 2 N. — 4. ipse igitur quadratus erit 4 Q. + 16. — 16 N. hæc æquabuntur vnitatibus 16 — 1 Q. Communis adiiciatur vtrimque defectus, & à similibus auferantur similia, fient 5 Q. æquales 16 N. & fit 1 N. $\frac{16}{5}$. Erit igitur alter quadratorum $\frac{256}{25}$. alter verò $\frac{144}{25}$ & vtriusque summa est $\frac{400}{25}$ seu 16. & vterque quadratus est.

Τὸν δοθέντα τετράγωνον διελεῖν εἰς δύο τετραγώνους. ἐπιτετάχθω δὴ τὸν ις̄ διελεῖν εἰς δύο τετραγώνους. καὶ τετάχθω ὁ πρῶτος δυναμεὶς μονάς. δεήσει ἄρα μονάδας ις̄ λεῖψιν δυνάμεως μονάδος ἴσας εἶναι τετραγώνῳ. πλάσω τὸ τετράγωνον ἀπὸ ζ̄. ἔσται δὴ πρώτη λεῖψις πόσων μ̄ ἔσων ἐξίσῃ ἃ τὸν ις̄ μ̄ πλάσεως. ἴσται ξ̄ β̄ λείψει μ̄ δ̄. αὐτὸς ἄρα ὁ πλευρὰ ἔσται δυνάμεων δ̄ μ̄ τῆς λείψεως ξ̄ ις̄. ταῦτα ἴσα μονάσι ις̄ λείψει δυνάμεως μιᾷ. κοινὴ προσκείσθω ἡ λεῖψις. καὶ ἀπὸ ὁμοίων ὅμοια. δυνάμεις ἄρα ε̄ ἴσαι ἀριθμοῖς κ̄ε̄. καὶ γίνεται ὁ ἀριθμὸς ις̄. πέμπτων. ἔσται ὁ μεῖζων σπ̄ εἰκοσιπέμπτων. ὁ δὲ μείζων εἰκοσιπέμπτων, ἐστὶ δὲ καὶ δύο συνθέντες πάλιν ὑ εἰκοσιπέμπτα, ἴσον μονάδας ις̄. καὶ ἔστιν ἑκάτερος τετράγωνος.

OBSERVATIO DOMINI PETRI DE FERMAT.

CVbum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos & generaliter nullam in infinitum vltra quadratum potestatem in duos eiusdem nominis fas est dividere cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

Figura 29. Nota al margen de Fermat, publicada en la edición que hizo su hijo de la Arithmetica de Diofanto.

No se ha conservado la copia personal de Fermat de la *Arithmetica*, pero él debe haber escrito en ella su nota al margen porque así lo dice Samuel. Es poco probable que Fermat haya dejado semejante tesoro sin abrir durante mucho tiempo, y su conjetura es tan natural que con toda probabilidad se le ocurrió en cuanto leyó la cuestión VIII del libro II. Evidentemente se preguntaría si podía conseguirse algo similar utilizando cubos en lugar de cuadrados, una pregunta natural para que se plantee un matemático. No encontró ningún ejemplo —podemos estar seguros de eso puesto que no existe ninguno— y tampoco tuvo éxito cuando ensayó potencias superiores, por ejemplo potencias cuartas. Decidió que

estas preguntas no tenían solución. Su nota al margen lo dice; se traduce así:

Es imposible dividir un cubo en dos cubos, o una potencia cuarta en dos potencias cuartas, o en general, cualquier potencia superior a la segunda en dos potencias similares. He descubierto una demostración verdaderamente maravillosa de esto, que no cabe en este estrecho margen.

En lenguaje algebraico, Fermat pretendía haber demostrado que la ecuación diofántica

$$x^n + y^n = z^n$$

no tiene soluciones en números naturales si n es cualquier entero mayor o igual que 3. Es obvio que estaba ignorando soluciones triviales en las que x o y es cero. Para no repetir la fórmula continuamente, me referiré a ella como la ecuación de Fermat.

Si Fermat en realidad tenía una demostración, nadie la ha encontrado. Al final se demostró que el teorema era cierto en 1995, más de tres siglos y medio después de que él lo enunciara por primera vez, pero los métodos van más allá de lo que estaba disponible en su época o de lo que él pudiera haber inventado. La búsqueda de una demostración tuvo una enorme influencia en el desarrollo de las matemáticas. Prácticamente dio lugar a la creación de la teoría de números algebraica, que floreció en el siglo XIX debido a un intento fallido de demostrar el teorema y una idea

brillante que lo salvó en parte. En los siglos XX y XXI desencadenó una revolución.

Quienes primero trabajaron en el último teorema de Fermat trataron de descartar potencias una por una. La demostración general de Fermat, a la que se aludía en su margen, puede haber existido o no, pero sí sabemos cómo demostró el teorema para potencias cuartas. La herramienta principal es la receta de Euclides para las tripletas pitagóricas. La potencia cuarta de cualquier número es el cuadrado del cuadrado de dicho número, de modo que cualquier solución de la ecuación de Fermat para potencias cuartas es una tripleta pitagórica en la que los tres números son asimismo cuadrados. Esta condición extra puede introducirse en la receta de Euclides y, después de algunas maniobras ingeniosas, lo que emerge es *otra solución* de la ecuación de Fermat para potencias cuartas[52]. Tras una página de álgebra, el problema se reduce al mismo problema, lo que no parece un avance. Sin embargo, hay una reducción real: los números en la segunda solución son menores que los de la primera, e hipotética, solución. De forma crucial, si la primera solución no es trivial —si x e y son distintos de cero— entonces lo mismo es cierto de la segunda solución. Fermat señaló que repetir este procedimiento llevaría a una secuencia de soluciones en la que los números se harían continuamente más pequeños. Sin embargo, cualquier secuencia decreciente de números naturales debe tener un final. Esto es una contradicción

[52] La demostración puede encontrase en muchos textos sobre teoría de números, por ejemplo Gareth A. Jones y J. Mary Jones, *Elementary Number Theory*, Springer, 1998, p. 227. En la web, véase http://en.wikipedia.org/wiki/infinite_descentXNon-solvability_og_r2_.2B_s4_.3D_t4.

lógica, de modo que la hipotética solución no existe. Él llamó a este método «descenso infinito». Ahora lo reconocemos como una demostración por inducción matemática, mencionada en el capítulo 4, y puede parafrasearse en términos de criminales mínimos. O, en este caso, modelos mínimos de virtud. Supongamos que existe un ciudadano virtuoso, una solución no trivial de la ecuación. Entonces existe un ciudadano virtuoso mínimo, una solución no trivial de la ecuación. Pero entonces el argumento de Fermat implica la existencia de un ciudadano mínimo aún menor: contradicción. Por consiguiente, no puede haber ciudadanos virtuosos. Desde entonces han estado apareciendo diferentes demostraciones para potencias cuartas, y ahora se conocen unas treinta.

Fermat explotó el simple hecho de que una potencia cuarta es un tipo especial de cuadrado. La misma idea muestra que para demostrar el último teorema de Fermat puede suponerse que la potencia n es 4 o un primo impar. Cualquier número n mayor que dos es divisible por 4 o por un primo impar p, de modo que toda potencia n-ésima es o una potencia cuarta o una potencia p-ésima. Durante los dos siglos siguientes, el último teorema de Fermat fue demostrado para exactamente tres primos impares: 3, 5 y 7. Euler trató con cubos en 1770; aunque hay una laguna en la demostración publicada, puede llenarse utilizando un resultado que Euler publicó en otro lugar. Legendre y Peter Lejeune-Dirichlet trataron con potencias quintas en torno a 1825. Gabriel Lamé demostró el último teorema de Fermat para potencias séptimas en 1839. Muchas demostraciones diferentes se encontraron más tarde

para estos casos. En el camino, varios matemáticos desarrollaron demostraciones cuando la potencia es 6, 10 y 14, pero estas fueron superadas por las demostraciones para 3, 5 y 7.

Cada demostración hace amplio uso de propiedades algebraicas que son especiales para la potencia concernida. No había ningún indicio de ninguna estructura general que pudiera demostrar el teorema para todas las potencias, o incluso para un número importante de potencias diferentes. Conforme las potencias se hacían más grandes, las demostraciones se hacían cada vez más complicadas. Se necesitaban ideas nuevas y había que abrir terreno nuevo. Sophie Germain, una de las grandes mujeres matemáticas, dividió el último teorema de Fermat para una potencia prima p en dos subcasos. En el primer caso, ninguno de los números x, y, z es divisible por p. En el segundo caso, uno de ellos lo es. Considerando primos «auxiliares» especiales relacionados con p, ella demostró que el primer caso del último teorema de Fermat no tiene soluciones para potencia prima impar menor que 100. Sin embargo, era difícil demostrar muchas cosas sobre primos auxiliares en general.

Germain mantenía correspondencia con Gauss, al principio utilizando un pseudónimo masculino, y él estaba muy impresionado por su originalidad. Cuando ella le reveló que era una mujer, él quedó aún más impresionado, y así lo dijo. A diferencia de muchos de sus contemporáneos, Gauss no suponía que las mujeres fueran incapaces de altos logros intelectuales, en particular de hacer investigación matemática. Más tarde Germain hizo un intento infructuoso de demostrar el primer caso del último teorema de

Fermat para todas las potencias pares, donde de nuevo es posible explotar la caracterización de Euclides de las tripletas pitagóricas. Guy Terjanian despachó finalmente las potencias pares en 1977. El segundo caso parecía un hueso mucho más difícil de roer, y nadie llegó muy lejos con él.

En 1847 Lamé, siguiendo su demostración para las potencias séptimas, tuvo una idea maravillosa. Requería la introducción de números complejos, pero en esa época ya todos se sentían cómodos con ellos. El ingrediente vital era el mismo que Gauss había explotado para construir un polígono regular de 17 lados (véase capítulo 3). Todo teórico de números lo conocía, pero hasta que llegó Lamé nadie se había preguntado en serio si podría ser precisamente la clave para demostrar el último teorema de Fermat.

En el sistema de los números reales, 1 tiene exactamente una raíz p-ésima (cuando p es impar), a saber, el propio 1. Pero en los números complejos, 1 tiene muchas raíces p-ésimas; de hecho, exactamente p de ellas. Esto es una consecuencia del teorema fundamental del álgebra, porque dichas raíces satisfacen la ecuación $x^p - 1 = 0$, que tiene grado p. Hay una bonita fórmula para estás raíces p-ésimas complejas de la unidad, como son llamadas, y muestra que son las potencias $1, \zeta, \zeta^2, \zeta^3 \ldots, \zeta^{p-1}$ de un número complejo concreto ζ[53]. La propiedad definitoria de estos números implica que $x^p + y^p$ se desdobla en p factores:

[53] Una raíz p-ésima de la unidad es el número complejo

$$\zeta = \cos 2\pi/p + i \operatorname{sen} 2\pi/p$$

$$x^p + y^p = (x + y)\,(x + \zeta y)\,(x + \zeta^2 y)\,\ldots\,(x + \zeta^{p-1} y)$$

Por la ecuación de Fermat, esta expresión es también igual a z^p, que es la potencia *p*-ésima de un entero. Ahora bien, es fácil ver que si un producto de números que no tienen ningún factor común es una potencia *p*-ésima, entonces cada número es por sí mismo una potencia *p*-ésima. Así, con unas pocas operaciones técnicas, Lamé

---

y las otras son sus potencias ζ^2, ζ^3... ζ^{p-1}. Para ver por qué, recordemos que las funciones trigonométricas seno y coseno se definen utilizando un triángulo rectángulo (véase Figura 56, *izquierda*). Para el ángulo *A*, utilizando las tradicionales *a, b, c* para los tres lados, definimos el seno (sen) y el coseno (cos) de *A* por

$$\text{sen } A = a/c \quad \cos A = b/c$$

Si hacemos $c = 1$ y situamos el triángulo en el plano complejo, como en la Figura 56 (*derecha*), el vértice en el que se encuentran *c* y *a* es el punto

$$\cos A + \text{i sen } A$$

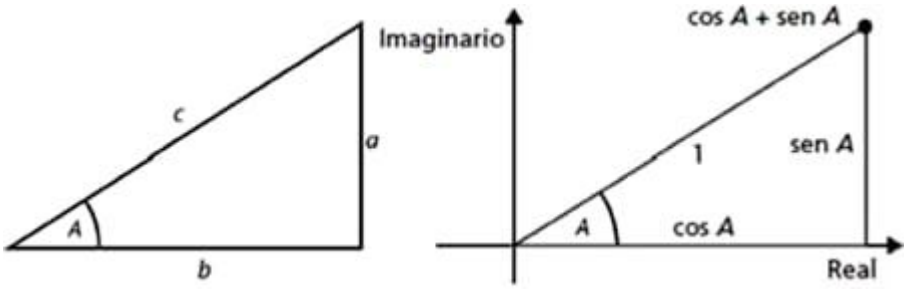


Figura 56. Izquierda: Definiendo el seno y el coseno. Derecha: Interpretación en el plano complejo.

Es ahora sencillo demostrar que para cualesquiera ángulos *A* y *B*,

$$(\cos A + \text{i sen } A)\,(\cos B + \text{i sen } B) = \cos (A + B) + \text{i sen } (A + B)$$

y esto lleva directamente a la fórmula de De Moivre

$$(\cos A + \text{i sen } A)^n = (\cos nA + \text{i sen } nA)$$

para cualquier entero positivo *n*. Por lo tanto

$$\zeta^p = (\cos 2\pi/p + \text{i sen } 2\pi/p)^p = \cos 2\pi + \text{i sen } 2\pi = 1$$

de modo que cada potencia 1, ζ, ζ^2, ζ^3..., ζ^{p-1} es una raíz *p*-ésima de la unidad. Nos detenemos aquí porque $\zeta^p = 1$, de modo que no aparecen números nuevos si tomamos potencias más altas.

pudo escribir cada factor como una potencia *p*-ésima. De esto él dedujo una contradicción.

Lamé anunció la demostración resultante del último teorema de Fermat a la Academia de París en marzo de 1847, dando el crédito de la idea básica a Joseph Liouville. Este se lo agradeció a Lamé, pero señaló un punto dudoso. No se puede dar por hecho el enunciado crucial que implica que cada factor es una potencia *p*-ésima. Depende de la unicidad de la factorización prima, no solo para enteros ordinarios, donde la propiedad es cierta, sino para los nuevos tipos de número que Lamé había introducido. Estas combinaciones de potencias de ζ se llaman enteros ciclotómicos. La palabra significa «que cortan el círculo», y se refiere a la conexión que Gauss había explotado. No es solo que la propiedad de factorización en primos única no estaba demostrada para enteros ciclotómicos, decía Liouville; es que incluso podría ser falsa.

Otros ya tenían dudas. Tres años antes, en una carta, Gotthold Eisenstein escribió:

> *Si tuviéramos un teorema que establece que el producto de dos números complejos solo puede ser divisible por un número primo cuando lo es uno de los factores —lo que parece completamente obvio— entonces tendríamos la teoría completa [de los números algebraicos] de una vez por todas; pero este teorema es totalmente falso.*

El teorema al que alude es el paso importante necesario para una demostración de la unicidad de la factorización prima. Eisenstein no

solo se estaba refiriendo a los números que necesitaba Lamé, sino también a números similares que aparecen en otras ecuaciones. Se denominan números algebraicos. Un número algebraico es un número complejo que satisface una ecuación polinómica con coeficientes enteros, siempre que el coeficiente que multiplica a la potencia más alta de x sea 1. Para cada uno de tales polinomios, obtenemos un campo de números algebraicos asociados (lo que significa que se pueden sumar, restar, multiplicar y dividir tales números para obtener números del mismo tipo) y su anillo (similar pero omitiendo la división) de enteros algebraicos. Estos son los objetos básicos que se estudian en la teoría de números algebraicos. Si, por ejemplo, el polinomio es $x^2 - 2$, entonces tiene solución $\sqrt{2}$. El campo consiste en todos los números $a + b\sqrt{2}$ con a, b racionales; el anillo de enteros consiste en los números de esta forma con a, b enteros. Una vez más pueden definirse factores primos, y son únicos. Hay algunas sorpresas: el polinomio $x^2 + x - 1$ tiene una solución $(\sqrt{5} - 1)/2$, de modo que a pesar de la fracción, este es un entero algebraico.

En la teoría de números algebraicos la dificultad no está en definir factores. Por ejemplo, un entero ciclotómico es un factor de (es decir, divide a) otro si el segundo es igual al primero multiplicado por algún entero ciclotómico. La dificultad tampoco está en definir primos: un entero ciclotómico es primo si no tiene factores, aparte de «unidades» triviales, que son los enteros ciclotómicos que dividen a 1. No hay problema en resolver un entero ciclotómico, o cualquier otro número algebraico, en factores primos. Simplemente hay que

seguir factorizandolo hasta que se agotan los factores. Hay una manera sencilla de demostrar que el procedimiento se detiene, y cuando lo hace cada factor debe ser primo. Entonces, ¿cuál es la dificultad? La unicidad. Si se ejecuta de nuevo el procedimiento, haciendo elecciones diferentes en el camino, se podría acabar con una lista diferente de factores primos.

A primera vista es difícil ver cómo puede suceder esto. Los factores primos son las piezas más pequeñas posibles en que puede dividirse el número. Es como tomar un juguete Lego y separarlo en los cubos que lo componen. Si hubiera otra manera de hacerlo, terminaría separando uno de dichos cubos en dos o más piezas. Pero entonces no sería un cubo. Por desgracia, la analogía con Lego es engañosa. Los números algebraicos no son así. Son más parecidos a cubos con enlaces móviles, capaces de acoplarse de maneras diferentes. Rompamos un cubo de una manera y las piezas resultantes encajan y ya no pueden separarse más. Rompámoslo de una manera diferente, y de nuevo las piezas resultantes encajan. Pero ahora son diferentes.

Le daré dos ejemplos. El primero utiliza solo enteros ordinarios; es fácil de entender pero tiene algunas características poco representativas. Luego le mostraré un ejemplo genuino.

Supongamos que viviéramos en un universo donde los únicos números que existían fueran 1, 5, 9, 13, 17, 21, 25 y así sucesivamente, números que en nuestro universo real tienen la forma $4k + 1$. Si se multiplican dos de estos números se obtiene otro número del mismo tipo. Definamos un número tal como primo si no

es el producto de dos números más pequeños *de ese tipo*. Por ejemplo, 25 no es primo porque es 5×5, y 5 es un número de la lista. Pero 21 *es* primo, en este nuevo sentido, porque sus factores ordinarios 3 y 7 no están en la lista. Son de la forma $4k + 3$, no $4k + 1$. Es fácil ver que todo número del tipo especificado es un producto de primos en el nuevo sentido. La razón es que los factores, si existen, deben ser menores. Finalmente, el proceso de factorización tiene que detenerse. Cuando lo hace, los factores concernidos son primos.

Sin embargo, este tipo de factorización en primos no es único. Consideremos el número 4389, que es 4×1097 + 1, de modo que es de la forma requerida. He aquí tres factorizaciones distintas en números de la forma requerida:

$$4389 = 21\times209 = 33\times133 = 57\times77$$

Yo afirmo que, con nuestra definición actual, todos estos factores son primos. Por ejemplo, 57 es primo, porque sus factores habituales 3 y 19 no son de la forma requerida. Lo mismo sucede con 21, 33, 77, 133 y 209. Ahora podemos explicar la falta de unicidad. En enteros ordinarios

$$4389 = 3\times7\times11\times19$$

y todos estos factores tienen la forma «errónea» $4k + 3$. Las tres factorizaciones en primos diferentes, en el nuevo sentido, aparecen agrupando estos números en pares:

$$(3\times7)\times(11\times19)\ (3\times11)\times(7\times19)\ (3\times19)\times(7\times11)$$

Necesitamos utilizar pares porque dos números de la forma $4k + 3$, multiplicados entre sí, dan un número de la forma $4k + 1$.

Este ejemplo muestra que el argumento «los factores deben ser únicos porque son las piezas más pequeñas» no funciona. Es cierto que hay piezas *más pequeñas* (21 = 3×7, por ejemplo) pero dichas piezas no están en el sistema concernido. La razón principal por la que este ejemplo no es muy representativo es que aunque multiplicar entre sí números de la forma $4k + 1$ da números de la misma forma, eso no es cierto para la suma. Por ejemplo, 5 + 5 = 10 no es de la forma requerida. Por eso, en la jerga del álgebra abstracta, no estamos trabajando en un anillo.

El segundo ejemplo no tiene este defecto, pero en compensación es algo más difícil de analizar. Es el anillo de los enteros algebraicos para el polinomio $x^2 - 15$. Este anillo consiste en todos los números $a + b\sqrt{15}$, donde a y b son enteros. En este, el número 10 tiene dos factorizaciones distintas:

$$10 = 2\times5 = (5 + \sqrt{15})\times(5 - \sqrt{15})$$

Puede demostrarse que los cuatro factores 2, 5, 5 + √15, 5 - √15 son primos[54].

Todo esto es mucho más claro ahora que lo era en 1847, pero no llevó mucho tiempo demostrar que las dudas de Liouville estaban justificadas. Dos semanas después de qué él las expresara, Wantzel informó a la academia de que la unicidad era cierta para algunos valores pequeños de *p*, pero su método de demostración fallaba para la potencia 23. Poco después, Liouville dijo a la academia que la factorización en primos única es *falsa* para enteros ciclotómicos correspondientes a $p = 23$. Ernst Kummer lo había descubierto tres años antes, pero no se lo había dicho a nadie porque estaba trabajando en un método para evitar el obstáculo. La demostración de Lamé funcionaba para valores menores de *p*, incluidos algunos nuevos: 11, 13, 17, 19. Pero para el caso general, la demostración estaba hecha jirones. Era una lección para no suponer que enunciados matemáticos plausibles son obvios. Puede que ni siquiera sean verdaderos.

---

[54] Introduzcamos la *norma*

$$N(a + b\sqrt{15}) = a^2 - 15b^2$$

que tiene la bonita propiedad

$$N(xy) = N(x)N(y)$$

Entonces

$$N(2) = 4\ N(5) = 25\ N(5 + \sqrt{15}) = 10\ N(5 - \sqrt{15}) = 10$$

Cualquier divisor propio de uno de estos cuatro números debe tener norma 2 o 5 (los divisores propios de sus normas). Pero las ecuaciones $a^2 - 15b^2 = 2$ y $a^2 - 15b^2 = 5$ no tienen soluciones enteras.
Por consiguiente, no existen divisores propios.

Kummer había estado pensando en el último teorema de Fermat, siguiendo líneas similares a la de Lamé. Él advirtió el obstáculo potencial, lo tomó en serio, lo investigó y descubrió que destruía esa aproximación. Encontró un ejemplo explícito de factorización en primos no única para enteros ciclotómicos basados en raíces 23 de la unidad. Pero Kummer no era alguien que abandonara con facilidad, y encontró una manera de evitar el obstáculo o, al menos, mitigar sus peores efectos. Su idea es especialmente transparente en el caso de los números de la forma $4k + 1$. La manera de recuperar la factorización única es incluir algunos números *nuevos*, que no están en el sistema en el que estamos interesados. Para ese ejemplo, lo que necesitamos son los números $4k + 3$ ausentes. O podemos llegar hasta el final e incluir también los enteros pares; entonces obtenemos los enteros, que son cerrados bajo la suma y la multiplicación. Es decir, si se suman o se multiplican dos enteros, el resultado es un entero.

Kummer dio con una versión de la misma idea. Por ejemplo, podemos recuperar la factorización en primos única en el anillo de todos los números $a + b\sqrt{15}$ incluyendo un nuevo número, a saber $\sqrt{15}$. Para obtener un anillo, resulta que también debemos incluir $\sqrt{3}$. Ahora

$$2 = (\sqrt{5} + \sqrt{3})\times(\sqrt{5} - \sqrt{3}) \qquad 5 = \sqrt{5}\times\sqrt{3}$$

y

$$5 + \sqrt{15} = \sqrt{5}\times(\sqrt{5} + \sqrt{3}) \qquad 5 - \sqrt{15} = \sqrt{5} \times(\sqrt{5} - \sqrt{3})$$

De modo que las dos factorizaciones aparecen al agrupar los cuatro números $\sqrt{5}$, $\sqrt{5}$, $\sqrt{5} + \sqrt{3}$, $\sqrt{5} - \sqrt{3}$ de dos maneras diferentes.

Kummer llamó a estos nuevos factores números ideales, porque en su formulación general no eran exactamente números. Eran símbolos que se comportaban de forma muy parecida a números. Él demostró que todo entero ciclotómico puede factorizarse unívocamente en números primos ideales. El truco era sutil: ni los enteros ciclotómicos ni los números ideales tenían factorización en primos única. Pero si se utilizaran los números ideales como ingredientes para factorización en primos de enteros ciclotómicos, el resultado era único.

Más tarde Richard Dedekind encontró una interpretación más refinada del procedimiento de Kummer, y esta es la que utilizamos ahora. A cada número ideal fuera del anillo concernido asoció un *conjunto* de números dentro del anillo. Llamó a dicho conjunto un ideal. Todo número en el anillo define un ideal: consiste en todos los múltiplos de dicho número. Si la factorización en primos es única, todo ideal es así. Cuando no lo es, hay ideales extra. Podemos definir el producto y la suma de ideales, e ideales primos, y Dedekind demostró que la factorización de *ideales* es única para todos los anillos de enteros algebraicos. Esto sugiere que para la mayoría de los problemas habría que trabajar con ideales, y no con los propios números algebraicos. Por supuesto, eso introduce

nuevas complejidades, pero la alternativa es normalmente quedarse atascado.

Kummer fue capaz de trabajar con sus números ideales, y lo hizo suficientemente bien para demostrar una versión del último teorema de Fermat con algunas hipótesis extra. Pero otros mortales encontraban los números ideales bastante difíciles, si no un poco místicos. Sin embargo, una vez vistos a la manera de Dedekind, los números ideales tenían perfecto sentido y con ellos despegó la teoría de números algebraicos. Una idea importante que salió de ella era una manera de cuantificar en qué medida falla la factorización única en un anillo de enteros algebraicos. A cada uno de tales anillos corresponde un número entero llamado su *número de clase*. Si el número de clase es 1, la factorización en primos es única; en los demás casos no lo es. Cuanto mayor es el número de clase, «menos única» es la factorización en primos, en un sentido significativo.

Ser capaz de cuantificar la falta de unicidad fue un gran paso adelante, y con esfuerzo extra salvó la estrategia de Lamé… a veces. En 1850 Kummer anunció que podía demostrar el último teorema de Fermat para muchos primos grandes, los que él llamaba regulares. Entre los primos menores que 100, solo 37, 59 y 67 son irregulares. Para todos los demás primos hasta dicho límite, y muchos por encima de él, sus métodos demostraban el último teorema de Fermat. La definición de un primo regular requiere el número de clase: un primo es regular si no divide el número de clase del anillo correspondiente de enteros ciclotómicos. Así que

para un primo regular, aunque la factorización en primos no sea única, la forma en que deje de serlo no implica al primo concernido de una manera esencial.

Kummer afirmaba que existen infinitos primos regulares, pero esta afirmación sigue sin estar demostrada. Irónicamente, en 1915, K. L. Jensen demostró que existen infinitos primos irregulares. Un extraño criterio para que un primo sea regular emergía de conexiones con el análisis. Incluye una secuencia de números descubiertos independientemente por el matemático japonés Seki Takazu (o Kōwa) y el matemático suizo Jacob Bernoulli, llamados números de Bernoulli. Este criterio muestra que los diez primeros primos irregulares son 37, 59, 67, 101, 103, 131, 149, 157, 233 y 257. Profundizando más en la estructura de los números ciclotómicos, Dmitri Mirimanoff despachó el primer primo irregular, 37, en 1893. Para 1905 él había demostrado el último teorema de Fermat hasta $p = 257$. Harry Vandiver desarrolló algoritmos para ordenador que ampliaron ese límite. Utilizando estos métodos, John Selfridge y Barry Pollack demostraron el teorema hasta la potencia 25.000 en 1967, y S. Wagstaff lo aumentó hasta 100.000 en 1976.

Las pruebas a favor de la verdad del último teorema de Fermat se estaban acumulando, pero la consecuencia más importante era que si el teorema fuera falso, entonces un contraejemplo —un ejemplo que mostrara su falsedad— sería tan gigantesco que nadie sería nunca capaz de encontrarlo. Otra consecuencia era que métodos como los de Kummer estaban tropezando con los mismos problemas que afectaban al trabajo de los primeros pioneros: potencias más

altas requerían un tratamiento especial y más complicado. De modo que esta línea de ataque llegó lentamente a un alto.

Cuando uno se queda atascado en un problema matemático, el consejo de Poincaré es claro: dejarlo y hacer otra cosa. Con suerte y viento favorable, ya surgirá una nueva idea. Los teóricos de números no siguieron de modo consciente su consejo, pero en cualquier caso hicieron lo que él había recomendado. Tal como predecía Poincaré, la táctica funcionó. Algunos teóricos de números dirigieron su atención a las curvas elípticas (véase capítulo 6). Irónicamente, esta área resultó tener un vínculo sorprendente e inesperado con el último teorema de Fermat, lo que llevó a la demostración de Wiles. Para describir este vínculo se requiere otro concepto adicional: el de una función modular. La discusión va a hacerse un poco técnica, pero hay una historia razonable detrás de las ideas y todo lo que vamos a necesitar es un esbozo general. Siga conmigo.

En el capítulo 6 vimos que la teoría de funciones elípticas tuvo un efecto profundo sobre el análisis complejo. En la década de 1830 Joseph Liouville descubrió que la variedad de funciones elípticas es bastante limitada. Dados los dos períodos, existe una función elíptica especial, la función de Weierstrass, y cualquier otra función elíptica con esos dos períodos es una simple variante. Esto implica que las únicas funciones doblemente periódicas que es necesario entender son las funciones de Weierstrass, una por cada par de períodos.

Desde el punto de vista geométrico, la estructura doblemente periódica de una función elíptica puede reformularse en términos de un retículo en el plano complejo: todas las combinaciones enteras $mu + nv$ de los dos períodos u y v, con enteros m y n (véase Figura 30). Si tomamos un número complejo z y le sumamos uno de estos puntos del retículo, la función elíptica en este nuevo punto tiene el mismo valor que tenía en el punto original. En otras palabras, la función elíptica tiene la misma simetría que el retículo.

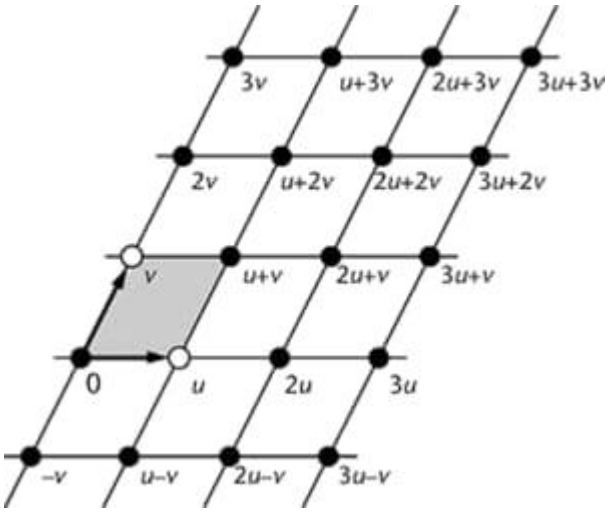


Figura 30. El retículo se forma a partir de todas las combinaciones enteras de los dos períodos.

Los analistas habían descubierto una fuente mucho más rica de simetrías del plano complejo, conocida como transformaciones de Möbius. Estas cambian z en $(az + b)/(cz + d)$, para constantes complejas a, b, c, d. Las simetrías del retículo son tipos especiales de transformaciones de Möbius, pero hay otras. Conjuntos de puntos análogos al retículo siguen existiendo en este escenario más

general. Un retículo define una pauta de teselación en el plano euclídeo: utilizar un paralelogramo como tesela y colocar sus vértices en los puntos del retículo (véanse Figuras 26 y 30). Utilizando transformaciones de Möbius, podemos construir pautas de teselación en una geometría no euclídea apropiada, el plano hiperbólico. Podemos identificar esta geometría con una región del plano complejo en la que las líneas rectas se reemplazan por arcos de círculo.

Existen pautas de teselación altamente simétricas en la geometría hiperbólica. Para cada una de ellas podemos construir funciones complejas que repiten los mismos valores en cada tesela. Se conocen como funciones modulares y son generalizaciones naturales de las funciones elípticas. La geometría hiperbólica es una disciplina muy rica, y el abanico de pautas de teselación es mucho más amplio que en el plano euclídeo. Por ello, el análisis complejo empezó reflexionando seriamente sobre la geometría no euclídea. Entonces apareció un vínculo profundo entre análisis y teoría de números. Las funciones modulares son a las curvas elípticas lo que las funciones trigonométricas son al círculo.

Recordemos que el círculo unidad consiste en los puntos (x, y) tales que $x^2 + y^2 = 1$. Supongamos que A es un número real, y hagamos

$x = \cos A$ $y = \text{sen } A$

Entonces la definición de seno y coseno nos dice que este punto se encuentra sobre el círculo unidad. Más aún, todo punto en el círculo unidad es de esta forma. En la jerga matemática, las funciones trigonométricas *parametrizan* el círculo. Algo muy

parecido sucede en el caso de las funciones modulares. Si definimos *x* e *y* utilizando funciones modulares adecuadas de un parámetro *A*, el punto correspondiente se encuentra en una curva elíptica —la misma curva elíptica, cualquiera que sea el valor que tome *A*—. Hay más formas abstractas de hacer preciso este enunciado, y quienes trabajan en el área las utilizan porque son más convenientes, pero esta versión resalta la analogía con la trigonometría y el círculo. Esta conexión genera una curva elíptica para cada función modular, y la variedad de funciones modulares es enorme: todas las teselaciones simétricas del plano hiperbólico. De modo que un tremendo número de curvas elípticas pueden relacionarse con funciones modulares. ¿Qué curvas elípticas pueden obtenerse de esta manera? Ese resultó ser el corazón de la materia.

Este «eslabón perdido» adquirió importancia por primera vez en 1975, cuando Yves Hellegouarch advirtió una curiosa conexión entre el último teorema de Fermat y las curvas elípticas. Gerhard Frey desarrolló más la idea en dos artículos publicados en 1982 y 1986. Supongamos, como siempre, que *p* es un primo impar. Supongamos también, con la esperanza de llegar a una contradicción, que existen enteros *a, b* y *c* distintos de cero que satisfacen la ecuación de Fermat, de modo que $a^p + b^p = c^p$. Ahora sacamos el conejo de la chistera con un gesto teatral: consideramos la curva elíptica

$$y^2 = x\,(x - a^p)(x - b^p)$$

Esta se denomina curva elíptica de Frey. Frey aplicó a esta la maquinaria de las curvas elípticas y lo que emergió era una cadena de coincidencias aún más extrañas. Su hipotética curva elíptica es en verdad muy extraña. Parece que no tiene sentido. Frey demostró que tiene tan poco sentido que no puede existir. Y eso, por supuesto, demuestra el último teorema de Fermat al proporcionar la contradicción requerida.

Sin embargo, había una laguna, y Frey era perfectamente consciente de ello. Para probar que esta hipotética curva elíptica no existe hay que demostrar que si existiera sería modular, es decir, sería una de las curvas que surgen a partir de funciones modulares. Acabamos de ver que tales curvas son habituales, y en esa época nadie había encontrado nunca una curva elíptica que *no* fuera modular. Parecía probable que la curva de Frey debía ser modular, pero era una curva hipotética, los números *a, b, c* no eran conocidos, y si la curva *fuera* modular, entonces no existiría siquiera. Sin embargo, había una manera de tratar todos estos puntos: demostrar que *toda* curva elíptica es modular. Entonces la curva de Frey, hipotética o no, tendría que ser modular si existiera. Y si no existía, la demostración era completa en cualquier caso.

La afirmación de que toda curva elíptica es modular se denomina conjetura de Taniyama-Shimura. Recibe este nombre por dos matemáticos japoneses, Yutaka Taniyama y Goro Shimura. Ambos se conocieron de un modo casual, al tratar de sacar prestado el mismo libro de la biblioteca al mismo tiempo y por la misma razón. Esto desencadenó una larga colaboración. En 1955 Taniyama

estaba en una conferencia de matemáticas en Tokio y los participantes más jóvenes fueron invitados a confeccionar una lista de preguntas abiertas. Taniyama aportó cuatro, todas las cuales incidían en una relación entre funciones modulares y curvas elípticas. Él había calculado algunos números asociados con una función modular concreta y advirtió que los mismos números exactamente aparecían en conexión con una curva elíptica concreta. Una coincidencia así suele ser signo de que no es en realidad una coincidencia, de que debe haber una explicación razonable para ello. Ahora se sabe que la igualdad de estos números equivale a que la curva elíptica sea modular; de hecho, esa es la definición preferida en la literatura de investigación. En cualquier caso, Taniyama estaba tan intrigado que calculó los números para otras pocas funciones modulares y encontró que también estas corresponden a curvas elípticas específicas.

Empezó a preguntarse si algo similar sería válido para toda curva elíptica. La mayoría de quienes trabajan en el campo consideraban esto demasiado bueno para ser cierto, una quimera para la que la evidencia era escasa. Shimura fue uno de los pocos que pensaba que la conjetura era digna de estudio. Pero Shimura pasó el curso 1957-1958 como profesor visitante en Princeton, y mientras él estaba en el extranjero Taniyama se suicidó. Dejó una nota que decía, en parte: «En cuanto a la causa de mi suicidio, yo mismo no la entiendo muy bien, pero no es el resultado de un incidente concreto, ni de una materia específica. Solo puedo decir que me encuentro en un estado mental en que he perdido confianza en mi

futuro». En esa época había pensado en contraer matrimonio, y su prometida, Misako Suzuki, se suicidó un mes después. Su nota de suicidio incluía: «Ahora que se ha ido, yo debo hacerlo también para unirme con él».

Shimura continuó trabajando en la conjetura, y a medida que acumulaba pruebas a su favor empezó a pensar que realmente podía ser cierta. La mayoría de los que trabajaban en el área discrepaban. Simon Singh[55] menciona una entrevista con Shimura en la que este recordaba cómo trataba de explicarla a uno de sus colegas:

> *El profesor dijo: «He oído que usted propone que algunas ecuaciones elípticas pueden estar relacionadas con formas modulares».*

«¡No, usted no lo entiende! —respondió Shimura—. No se trata solamente de algunas ecuaciones elípticas; ¡son *todas* las ecuaciones elípticas!».

Pese a este escepticismo, Shimura perseveró, y al cabo de muchos años la propuesta se había hecho suficientemente respetable para ser conocida como la conjetura de Taniyama-Shimura. Entonces André Weil, uno de los grandes teóricos de números del siglo XX, encontró otras muchas pruebas a favor de la conjetura, las publicó y expresó la creencia de que muy bien podría ser cierta. Se hizo conocida como la conjetura de Taniyama-Shimura-Weil. No hay un nombre definitivo: se han utilizado todo tipo de permutaciones de

[55] Simon Singh, *Fermat's Last Theorem*, Fourth Estate, 1997.

subconjuntos de los tres matemáticos. Yo me atendré a «conjetura de Taniyama-Shimura».

En los años sesenta del siglo pasado otro peso pesado, Robert Langlands, se dio cuenta de que la conjetura de Taniyama-Shimura podía verse como un elemento en un programa mucho más amplio y ambicioso que unificaría las teorías de números algebraica y analítica. Formuló todo un paquete de conjeturas relacionadas con esta idea, ahora conocidas como el programa de Langlands. Era incluso más especulativo que la conjetura de Taniyama-Shimura, pero tenía una elegancia irresistible; era el tipo de matemáticas que deberían ser verdaderas por su belleza. A lo largo de los años setenta el mundo matemático se fue acostumbrando a la belleza salvaje del programa de Langlands, que empezó a ser aceptado como uno de los objetivos centrales de la teoría de números algebraica. El programa de Langlands parecía ser el camino correcto para avanzar, solo con que alguien pudiera dar el primer paso.

En este momento Frey advirtió que la aplicación de la conjetura de Taniyama-Shimura a su curva elíptica demostraría el último teorema de Fermat. Sin embargo, por entonces había surgido otro problema con la idea de Frey. Cuando él dio una charla sobre la misma en 1984, la audiencia detectó una laguna en su argumento clave: la curva es tan extraña que no puede ser modular. Jean-Pierre Serre, una de las figuras destacadas en el área, llenó rápidamente la laguna, pero tuvo que invocar otro resultado que también carecía de demostración, la conjetura de reducción de nivel especial. No obstante, esta conjetura fue demostrada por Ken Ribet

en 1986. Ahora el único obstáculo para una demostración del último teorema de Fermat era la conjetura de Taniyama-Shimura, y el consenso empezó a cambiar. Serre predijo que el último teorema de Fermat sin duda sería resuelto en menos de una década. Cómo lo sería exactamente era otra cuestión, pero una sensación general flotaba en el ambiente: las técnicas relacionadas con funciones modulares se estaban haciendo tan potentes que pronto alguien haría que funcionase la aproximación de Frey.

Ese alguien fue Andrew Wiles. En un programa de televisión dedicado a su demostración, dijo:

> *Yo tenía diez años y... encontré un libro de matemáticas que contaba la historia de este problema [último teorema de Fermat]; que alguien había [planteado] este problema hace trescientos años pero nadie había visto una demostración, nadie sabía si había una demostración, y desde entonces se ha buscado la demostración. Y había un problema que yo, con diez años, podía entender, pero ninguno de los grandes matemáticos del pasado había sido capaz de resolver. Y desde ese momento, por supuesto, yo mismo traté de resolverlo. Tanto era el reto y tan bello era el problema.*

En 1971 Wiles obtuvo un grado en matemáticas en Oxford y se trasladó a Cambridge para hacer su doctorado. Su tutor, John Coates, le advirtió (correctamente) que el último teorema de Fermat era demasiado difícil para una tesis. Así que en lugar de ello, Wiles

se puso a trabajar en funciones elípticas, entonces considerada un área de investigación mucho más prometedora. En 1985-1986 estaba en París en el Institut des Hautes Études Scientifiques (Instituto de Estudios Científicos Avanzados), uno de los más importantes institutos de investigación matemática del mundo. Muchos de los mejores investigadores pasan por allí en algún momento; si uno es un matemático, es un gran lugar para pasar un tiempo. Entre los visitantes estaba Ribet, y su demostración de la conjetura de reducción del nivel especial electrizó a Wiles. Ahora él podía hacer una investigación completamente respetable en curvas elípticas, tratando de demostrar la conjetura de Taniyama-Shimura, y al mismo tiempo podía tratar de cumplir el sueño de su infancia de demostrar el último teorema de Fermat.

Puesto que todo el mundo en el área conocía la conexión, había un motivo de preocupación. Supongamos que Wiles consiguiera juntar una demostración casi completa, con algunas pequeñas lagunas que necesitaran trabajo extra. Supongamos que algún otro lo supiera y llenara la laguna. Entonces, desde el punto de vista técnico, esta persona sería la que había demostrado el último teorema de Fermat. En general los matemáticos no se comportan así, pero cuando el premio es tan grande es prudente tomar precauciones. Por eso Wiles llevó a cabo su investigación en secreto, algo que los matemáticos no suelen hacer. No es que él no confiara en sus colegas. Es simplemente que no podía correr el más mínimo riesgo de ser adelantado en la línea de llegada.

Trabajó durante siete años, encerrado en el ático de su casa donde había un despacho. Solo su mujer y el director de su departamento sabían en qué estaba trabajando. En paz y reclusión atacó el problema con todas las técnicas que pudo aprender, hasta que los muros del castillo empezaron a temblar bajo el asalto. En 1991 Coates le pasó algunos nuevos resultados demostrados por Mattheus Flach. La grieta en el muro empezó a ensancharse cada vez más deprisa a medida que continuaba el asedio.

En 1993 la demostración estaba completa. Ahora había que revelarla al mundo. Aún cauto, Wiles no quería arriesgarse a hacer pública su solución por si salía a la superficie algún error, algo que le sucedió a Yoichi Miyaoka en 1988, cuya pretensión de una demostración salió en los medios de comunicación solo para que se encontrara un error fatal. Por ello Wiles decidió impartir una serie de tres conferencias en el Instituto Isaac Newton en Cambridge, un centro de investigación internacional de matemáticas recientemente fundado. El título era técnico e inocuo: «Formas modulares, curvas elípticas y teoría de Galois». No engañó a muchos: todos sabían que Wiles iba tras algo grande.

En la tercera conferencia, Wiles esbozó una demostración de un caso especial de la conjetura de Taniyama-Shimura. Había descubierto que también funcionaría algo menos ambicioso: demostrar que la curva de Frey, si existe, debe pertenecer a una clase especial de curvas elípticas, las «semiestables», y demostrar asimismo que todas las curvas en dicha clase deben ser modulares. Wiles demostró entonces ambos resultados. Al final de la

conferencia escribió en la pizarra un corolario —un teorema complementario que se sigue directamente de lo que se acaba de demostrar—. El corolario era el último teorema de Fermat.

Cuando Shimura supo del anuncio de Wiles, su comentario fue breve y directo: «Yo lo dije».

Aquí podría haber acabado todo. Pero aguardaba un giro del destino. La demostración aún tenía que ser evaluada por expertos, y, como suele suceder, ese proceso puso de manifiesto algunos puntos que necesitaban más explicación. Wiles respondió a la mayor parte de estos comentarios, pero uno de ellos necesitaba un replanteamiento. Avanzado 1993 comunicó que retiraba su pretensión hasta que pudiera llenar una laguna lógica que había surgido. Pero ahora estaba obligado a actuar a plena luz pública, exactamente lo que había intentado evitar.

En marzo de 1994 no había aparecido ninguna demostración corregida, y Faltings expresó una opinión ampliamente extendida en la comunidad matemática: «Si [reparar la demostración] fuera fácil, él ya lo hubiera resuelto. Estrictamente hablando, no era una demostración cuando fue anunciada». Weil comentó: «Creo que él tiene algunas buenas ideas… pero la demostración no existe… lo que demuestra que el último teorema de Fermat es como escalar el Everest. Si un hombre quiere escalar el Everest y se queda a cien metros de la cima, no ha escalado el Everest». Todos podían conjeturar cómo iba a terminar. Lo habían visto antes. La demostración se había venido abajo, tendría que ser retirada y el último teorema de Fermat seguiría vivo y coleando.

Wiles se negó a conceder la derrota, y su antiguo estudiante Richard Taylor se unió a la búsqueda. Ahora estaba clara la raíz de la dificultad: los resultados de Flach no eran muy adecuados para la tarea. Ellos trataron de modificar los métodos de Flach, pero nada parecía funcionar. Luego, en un destello de inspiración, Wiles comprendió de repente cuál era el obstáculo. «Vi que lo que había impedido que [el método de Flach] funcionara era algo que haría que funcionara otro método que yo había ensayado previamente». Era como si los soldados que asediaban el castillo se hubieran dado cuenta de que su ariete nunca funcionaría porque los defensores seguían arrojando piedras sobre el mismo, pero esas mismas piedras podían ser cargadas en una catapulta y utilizadas para romper la puerta.

Para abril de 1995 la nueva demostración estaba acabada, y esta vez no había lagunas ni errores. Rápidamente siguió la publicación, dos artículos en los ultraprestigiosos *Annals of Mathematics*. Wiles se convirtió en una celebridad internacional, recibió varios premios importantes, fue nombrado caballero… y volvió a su investigación, con la misma pasión que antes.

Lo realmente importante de la solución de Wiles no es, ni mucho menos, el último teorema de Fermat. Como he dicho, nada vital depende de la respuesta. Si alguien hubiera encontrado tres números de cien dígitos y un primo de doscientos cincuenta dígitos que proporcionaran un contraejemplo a la afirmación de Fermat, entonces el teorema habría sido falso, pero ninguna área crucial de las matemáticas se hubiera visto reducida. Por supuesto, un ataque

directo por ordenador no sería capaz de buscar en números tan grandes, de modo que habría que ser sorprendentemente inteligente para establecer algo semejante, pero un resultado negativo no habría causado ningún infarto.

La importancia real de la solución reside en la demostración del caso semiestable de la conjetura de Taniyama-Shimura. En menos de seis años Christophe Breuil, Brian Conrad, Fred Diamond y Taylor ampliaron los métodos de Wiles para tratar no solo el caso semiestable sino todas las curvas elípticas. Ellos demostraron la conjetura de Taniyama-Shimura completa, y la teoría de números ya nunca sería la misma. A partir de entonces, cuando quiera que alguien encontrara una curva elíptica, estaba garantizado que era modular, de modo que se abriría un montón de métodos analíticos. Dichos métodos ya se han utilizado para resolver otros problemas en la teoría de números, y más se mostrarán en el futuro.

Capítulo 8

Caos orbital

El problema de los tres cuerpos

Un viejo chiste afirma que se puede decir hasta qué punto es avanzada una teoría física por el número de cuerpos interactuantes que no puede manejar. La ley de la gravedad de Newton entra en problemas con tres cuerpos. La relatividad general tiene dificultades al tratar con dos cuerpos. La teoría cuántica se queda corta para un cuerpo, y la teoría cuántica de campos entra en problemas cuando no hay *ningún* cuerpo, en el vacío. Como muchos chistes, este contiene un grano de verdad[56]. En particular, la interacción gravitatoria de tan solo tres cuerpos, que se supone, obedecen a la ley de la inversa del cuadrado de Newton, confundió al mundo matemático durante siglos. Aún lo hace, si lo que uno quiere es una bonita fórmula para las órbitas de dichos cuerpos. De hecho, ahora sabemos que la dinámica de tres cuerpos es caótica: tan irregular que tiene elementos de aleatoriedad.

Todo esto está en tremendo contraste con el sorprendente éxito de la teoría gravitatoria newtoniana, que explicaba, entre muchas otras cosas, la órbita de un planeta alrededor del Sol. La respuesta es la que Kepler ya había deducido empíricamente a partir de las observaciones astronómicas de Marte: una elipse. Aquí solo hay dos cuerpos: Sol y planeta. El siguiente paso obvio es utilizar la ley de la

[56] O quizá no. Vladimir Krivchenkov ha señalado que la energía del estado fundamental y los primeros estados excitados para el problema de los tres cuerpos cuántico pueden calcularse a mano. Pero en mecánica clásica, el problema análogo es menos tratable debido al caos.

gravedad de Newton para escribir la ecuación para las órbitas de tres cuerpos, y resolverla. Pero no hay ninguna caracterización geométrica clara de las órbitas de tres cuerpos, ni siquiera una fórmula en geometría de coordenadas. Hasta finales del siglo XIX muy poco se sabía sobre el movimiento de tres cuerpos celestes, incluso si uno de ellos fuera tan minúsculo que su masa podía ignorarse.

Nuestro conocimiento de la dinámica de tres (o más) cuerpos ha aumentado espectacularmente desde entonces. Una gran parte de ese progreso ha sido una comprensión cada vez mayor de lo difícil que es la pregunta, y por qué. Eso puede parecer un paso atrás, pero a veces la mejor manera de avanzar es hacer una retirada estratégica y ensayar alguna otra cosa. En el caso del problema de los tres cuerpos este plan de campaña ha conseguido algunos éxitos reales, cuando un ataque frontal se hubiera empantanado sin remedio.

Los primeros seres humanos no pueden haber dejado de advertir que la Luna se mueve lentamente a través del cielo nocturno con respecto al fondo estrellado. Las estrellas también parecen moverse, pero lo hacen como un todo, como minúsculos alfileres de luz en una enorme bóveda giratoria. La Luna es claramente especial en otro aspecto: es un gran disco brillante que cambia de forma desde luna nueva hasta luna llena y vuelta a empezar. No es un alfiler de luz como una estrella.

Algunos de estos alfileres de luz también desobedecen las reglas. Se pasean. No cambian de posición con respecto a las estrellas tan

rápidamente como la Luna pero, incluso así, no hace falta observar el cielo durante muchas noches para ver que algunos se están moviendo. Cinco de estas estrellas errantes son visibles a simple vista; los griegos las llamaban *planētēs* («vagabundos»). Son, por supuesto, los planetas, y los cinco que han sido reconocidos desde tiempos antiguos son los que ahora llamamos Mercurio, Venus, Marte, Júpiter y Saturno —todos nombres de dioses romanos—. Con la ayuda de telescopios ahora conocemos dos más: Urano y Neptuno. Más nuestra propia Tierra, por supuesto. Plutón ya no cuenta como planeta, gracias a una controvertida decisión sobre terminología que tomó la Unión Astronómica Internacional en 2006.

Conforme los antiguos filósofos, astrónomos y matemáticos estudiaban los cielos, se dieron cuenta de que los planetas no vagan al azar. Siguen órbitas enrevesadas pero bastante predecibles, y vuelven prácticamente a la misma posición en el cielo nocturno a intervalos de tiempo muy regulares. Ahora explicamos estas pautas como movimiento periódico en una órbita cerrada, con una pequeña contribución del propio movimiento orbital de la Tierra. También reconocemos que la periodicidad no es exacta aunque está muy próxima a ello. Mercurio tarda casi ochenta y ocho días en dar una vuelta alrededor del Sol, mientras que Júpiter tarda casi doce años. Cuanto más lejos del Sol está el planeta, más tiempo tarda en completar una órbita.

El primer modelo cuantitativamente preciso del movimiento de los planetas fue el sistema de Ptolomeo, que debe su nombre a Claudio Ptolomeo, quien lo describió en su *Almagesto* («El [tratado] máximo»)

aproximadamente en 150 d. C. Es un modelo geocéntrico (centrado en la Tierra) en el que todos los cuerpos celestes orbitan en torno a la Tierra. Se mueven como si estuvieran engarzados en una serie de esferas gigantescas, cada una de las cuales rota a una velocidad fija alrededor de un eje que a su vez puede estar engarzado en otra esfera. Se necesitaban combinaciones de muchas esferas rotatorias para representar el complejo movimiento de los planetas en términos del ideal cósmico de rotación uniforme en un círculo, el ecuador de la esfera. Con suficientes esferas y las elecciones correctas de sus ejes y velocidades, el modelo se corresponde muy estrechamente con la realidad.

Nicolás Copérnico modificó el esquema de Ptolomeo en varios aspectos. El más radical fue hacer que todos los cuerpos, salvo la Luna, den vueltas alrededor del Sol, lo que simplificaba de modo considerable la descripción. Era un modelo heliocéntrico. Esta propuesta contradecía a la iglesia Católica, pero finalmente la visión científica prevaleció y las personas instruidas aceptaron que la Tierra daba vueltas alrededor del Sol. En 1596 Kepler defendía el sistema copernicano en su *Mysterium cosmographicum* («El misterio cosmográfico»), cuyo punto culminante era su descubrimiento de una relación matemática entre la distancia de un planeta al Sol y su período orbital. Según esta, si nos movemos hacia fuera a partir del Sol la razón del incremento en período de un planeta al siguiente es dos veces la diferencia de los radios orbitales. Más tarde decidió que esta relación era demasiado imprecisa para ser correcta, pero sembró las semillas de una relación más exacta en su trabajo

futuro. Kepler también explicó el espaciado de los planetas en términos de los cinco sólidos regulares, limpiamente anidados unos dentro de otros, separados por las esferas que los sostenían. Cinco sólidos explicaban por qué había cinco planetas, pero ahora reconocemos ocho, de modo que esta propiedad ya no es una ventaja. Hay 120 maneras diferentes de ordenar cinco sólidos, y es probable que una de estas se aproxime a las proporciones celestes dadas por las órbitas planetarias. De modo que es simplemente una aproximación accidental, que encaja con calzador en la naturaleza una pauta sin significado.

En 1600 el astrónomo Tycho Brahe contrató a Kepler para que le ayudara a analizar sus observaciones, pero se interpusieron problemas políticos. Tras la muerte de Brahe, Kepler fue nombrado matemático imperial por Rodolfo II. En su tiempo libre trabajaba en las observaciones de Marte que había hecho Brahe. Un resultado fue *Astronomia nova* («Una nueva astronomía») de 1609, donde presentaba otras dos leyes del movimiento planetario. La primera ley de Kepler afirma que los planetas se mueven en elipses. Él lo había establecido para Marte, y parecía probable que lo mismo fuera cierto para los otros planetas. Al principio supuso que una forma ovalada ajustaría los datos, pero eso no funcionó, de modo que probó con una elipse. Tampoco esto le pareció aceptable y buscó una descripción matemática diferente para la forma de la órbita. Finalmente se dio cuenta de que esta era en realidad solo otra manera de definir una elipse[57]:

[57] Citado en Arthur Koestler, *The Sleepwalkers*, Penguin Books, 1990, p. 338.

> *Dejé aparte [la nueva definición], y volví a las elipses, creyendo que esta era una hipótesis muy diferente, mientras que las dos, como demostraré en el próximo capítulo, son una y la misma... ¡Ah, qué insensato he sido!*

La segunda ley de Kepler afirma que el planeta barre áreas iguales en tiempos iguales. En 1619, en su *Harmonices mundi* («Armonías del mundo»), Kepler completó sus tres leyes con una relación mucho más precisa entre distancias y períodos: el cubo de la distancia (la mitad del eje mayor de la elipse) es proporcional al cuadrado del período.

El escenario estaba ahora preparado para Isaac Newton. En sus *Philosophiae naturalis principia mathematica* («Principios matemáticos de la filosofía natural») de 1687, Newton demostró que las tres leyes de Kepler son equivalentes a una única ley de la gravitación: dos cuerpos se atraen mutuamente con una fuerza que es proporcional a sus masas e inversamente proporcional al cuadrado de la distancia entre ellos. La ley de Newton tenía una gran ventaja: se aplicaba a cualquier sistema de cuerpos, por muchos que pudieran ser. El precio que había que pagar era la manera en que la ley determinaba las órbitas: no como formas geométricas, sino como soluciones de una ecuación diferencial que incluía las aceleraciones de los planetas. No está claro en absoluto cómo encontrar las formas de las órbitas planetarias, o las posiciones de los planetas en un instante dado, a partir de esta ecuación. Para ser franco, no está nada claro cómo encontrar sus

aceleraciones. De todas formas, la ecuación proporcionaba *implícitamente* dicha información. El problema estaba en hacerla explícita. Kepler ya lo había hecho para dos cuerpos, y la respuesta era órbitas elípticas con velocidades que barren áreas a un ritmo constante.

¿Y qué pasa con tres cuerpos?

Era una buena pregunta. De acuerdo con la ley de Newton todos los cuerpos en el Sistema Solar se influyen gravitatoriamente unos a otros. De hecho, todos los cuerpos en el universo entero se influyen gravitatoriamente unos a otros. Pero nadie en su sano juicio trataría de escribir ecuaciones diferenciales para todos los cuerpos en el universo. Como siempre, la manera de avanzar era simplificar el problema… pero no demasiado. Las estrellas están tan alejadas que su influencia gravitatoria sobre el Sistema Solar es despreciable a menos que se quiera describir como se mueve el Sol a medida que rota la galaxia. El movimiento de la Luna está influenciado sobre todo por otros dos cuerpos, la Tierra y el Sol, aparte de algunos efectos sutiles que implican a otros planetas. A principios del siglo XVIII esta cuestión escapó de los dominios de la astronomía y adquirió importancia práctica, cuando se entendió que el movimiento de la Luna podía ser útil en la navegación. (No había GPS en aquellos días; ni siquiera había cronómetros para medir la longitud geográfica). Pero este método requería predicciones más precisas que las que podían proporcionar las teorías existentes. El lugar obvio donde empezar era desarrollar las implicaciones de la ley de Newton para tres cuerpos, que para este fin podían tratarse

como masas puntuales porque los planetas son extraordinariamente pequeños comparados con las distancias entre ellos. Entonces se resolvían las ecuaciones diferenciales resultantes. Sin embargo, los trucos que llevaban desde dos cuerpos a las elipses fallaban cuando un cuerpo extra entraba en la mezcla. Algunos pasos preliminares funcionaban, pero luego el cálculo quedaba bloqueado. En 1747 Jean D'Alembert y Alexis Clairaut, amargos rivales, compitieron por un premio de la Academia de Ciencias de París sobre el «problema de los tres cuerpos», que ambos abordaron a través de una aproximación numérica. El problema de los tres cuerpos había adquirido su nombre, y pronto llegó a ser uno de los grandes enigmas de las matemáticas.

Algunos casos especiales podían resolverse. En 1767 Euler descubrió soluciones en las que los tres cuerpos se encontraban en una línea recta en rotación. En 1772 Lagrange encontró soluciones similares en donde los cuerpos forman un triángulo equilátero en rotación, que se expande y contrae. Ambas soluciones eran periódicas: los cuerpos repetían la misma secuencia de movimientos indefinidamente. Sin embargo, incluso con simplificaciones drásticas no se llegaba a producir algo más general. Se podía suponer que uno de los cuerpos tenía una masa despreciable, se podía suponer que los otros dos se movían en círculos perfectos en torno a su centro de masas, una versión conocida como el problema de los tres cuerpos «restringido»... y *aún así* no se podían resolver las ecuaciones exactamente.

En 1860 y 1867 el astrónomo y matemáticos Charles-Eugène Delaunay atacó el caso específico del sistema Sol-Tierra-Luna utilizando teoría de perturbaciones, que considera el efecto de la gravedad del Sol sobre la Luna como un cambio pequeño superpuesto al efecto de la Tierra, y derivó fórmulas aproximadas en forma de series: muchos términos sucesivos sumados. Publicó sus resultados en 1860 y 1867; cada volumen tenía novecientas páginas y consistía básicamente en fórmulas. A finales de los años setenta del siglo pasado sus cálculos fueron comprobados utilizando programas de álgebra simbólica, y solo se encontraron dos errores pequeños y poco importantes.

Fue un cálculo heroico, pero la serie se aproximaba a su valor límite demasiado lentamente para tener uso práctico. Sin embargo, animó a otros a buscar soluciones en forma de series que convergieran con mayor rapidez. También destapó un gran obstáculo técnico para todas las aproximaciones de este tipo, conocido como el problema de los denominadores pequeños. Algunos términos de las series son fracciones, y los denominadores (la parte inferior) se hacen muy pequeños si los cuerpos están casi en resonancia: un estado periódico en el que sus períodos son múltiplos racionales uno de otro. Por ejemplo, las tres lunas más interiores de Júpiter, a saber Io, Europa y Ganímedes, tienen períodos de revolución en torno al planeta de 1,77 días, 3,55 días y 7,15 días, en razones 1: 2: 4 casi exactas. Las resonancias seculares, relaciones racionales entre las velocidades a las que giran los ejes de dos órbitas casi elípticas, son

una molestia especial, porque el error probable al evaluar una fracción se hace muy grande cuando su denominador es pequeño.

Si el problema de los tres cuerpos era difícil, el problema de *n*-cuerpos —cualquier número de masas puntuales que se mueven bajo la acción de la gravedad newtoniana— era sin duda más difícil. Pero la naturaleza nos presenta un ejemplo importante: el Sistema Solar entero. Este contiene ocho planetas, varios planetas enanos como Plutón y miles de asteroides, muchos de ellos bastante grandes. Por no mencionar los satélites, algunos de los cuales —Titán, por ejemplo— son más grandes que el planeta Mercurio. De modo que el Sistema Solar es un problema de diez cuerpos, o un problema de veinte cuerpos, o un problema de diez mil cuerpos, dependiendo de cuántos detalles se quieran incluir.

Para predicciones a corto plazo las aproximaciones numéricas son efectivas, y en astronomía mil años es poco tiempo. Entender cómo evolucionará el Sistema Solar durante cientos de millones de años es otra cosa muy diferente. Y una gran pregunta depende de ese tipo de visión a largo plazo: la estabilidad del Sistema Solar. Los planetas parecen moverse en órbitas casi elípticas y relativamente estables. Estas órbitas cambian un poco cuando otros planetas las perturban, por lo que el período podría cambiar en una fracción de segundo o el tamaño de la elipse podría no ser exactamente constante. ¿Podemos estar seguros de que estos suaves empujones es todo lo que va a suceder en el futuro? ¿Es representativo de lo que sucedió en el pasado, sobre todo en las primeras fases del Sistema Solar? ¿Permanecerá estable el Sistema Solar o

colisionarán dos planetas? ¿Podría un planeta escapar a los confines distantes del universo?

El año 1889 se celebraba el sexagésimo cumpleaños de Oscar II, rey de Noruega y Suecia. Como parte de las celebraciones, el matemático noruego Gösta Mittag-Leffler convenció al rey para que convocara un premio por la solución del problema de *n* cuerpos. Esta se obtendría no por una fórmula exacta —para entonces estaba claro que esto era mucho pedir—, sino por algún tipo de serie convergente. Poincaré se interesó y decidió empezar por una versión muy simple: el problema de los tres cuerpos restringido, en el que uno de los cuerpos tiene una masa despreciable, como una minúscula partícula de polvo. Si se aplica la ley de Newton ingenuamente a dicha partícula, la fuerza ejercida sobre ella es el producto de las masas dividido por el cuadrado de la distancia, pero una de las masas es cero, de modo que el producto es cero. Esto no ayuda mucho, porque la partícula de polvo simplemente sigue su propio camino, desacoplada de los otros dos cuerpos. En su lugar, se establece el modelo de modo que la partícula de polvo siente el efecto de los otros dos cuerpos pero estos la ignoran por completo. De modo que las órbitas de los dos cuerpos masivos son circulares y se mueven a una velocidad fija. Toda la complejidad del movimiento se invierte en la partícula de polvo.

Poincaré no resolvió el problema que planteaba el rey Oscar. Eso era demasiado ambicioso. Pero sus métodos eran tan innovadores, y él hizo tantos progresos, que pese a todo se le concedió el premio. Su investigación ganadora del premio fue publicada en 1890 y sugería

que incluso el problema de los tres cuerpos restringido podría no tener el tipo de respuesta que se había estipulado. Poincaré dividió su análisis en varios casos diferentes, dependiendo de las características generales del movimiento. En la mayoría de ellos podían obtenerse perfectamente soluciones en forma de serie. Pero había un caso en el que la órbita de la partícula de polvo era en extremo complicada.

Poincaré dedujo esta complejidad inevitable a partir de algunas otras ideas que había desarrollado, lo que hacía posible describir soluciones de las ecuaciones diferenciales sin resolverlas realmente. Esta «teoría cualitativa de ecuaciones diferenciales» fue la semilla a partir de la que ha crecido la moderna dinámica no lineal. La idea básica era explorar la geometría de las soluciones; más en concreto su topología, un tema en el que Poincaré estaba también profundamente interesado (véase capítulo 10). En esta interpretación, las posiciones y velocidades de los cuerpos son coordenadas en un espacio multidimensional. Conforme pasa el tiempo cualquier estado inicial sigue una trayectoria curva a través de este espacio. La topología de esta trayectoria, o el sistema entero de todas las trayectorias posibles, nos dice muchas cosas útiles sobre las soluciones.

Una solución periódica, por ejemplo, es una trayectoria que se cierra sobre sí misma para formar un lazo. Conforme pasa el tiempo, el estado recorre el lazo una y otra vez, repitiendo el mismo comportamiento en un proceso sin fin. El sistema es entonces periódico. Poincaré sugería que una buena manera de detectar tales

lazos es colocar una superficie multidimensional de modo que sea atravesada por el lazo. Ahora la llamamos sección de Poincaré. Soluciones que parten de puntos en esta superficie pueden volver con el tiempo a la superficie; el propio lazo vuelve exactamente al mismo punto, y soluciones que pasan por puntos vecinos vuelven siempre a la sección al cabo de aproximadamente un período. De modo que una solución periódica puede ser interpretada como un punto fijo de la «aplicación de primer retorno», que nos dice lo que les sucede a puntos de la superficie cuando vuelven a ella por primera vez, si es que lo hacen. Puede que esto no parezca un gran avance, pero reduce la dimensión del espacio, el número de variables en el problema. Esto es casi siempre una buena cosa.

La gran idea de Poincaré empieza a mostrar su verdadero valor cuando pasamos al siguiente tipo más complicado de solución, una combinación de varios movimientos periódicos. A modo de ejemplo sencillo, la Tierra da una vuelta alrededor del Sol cada 365 días aproximadamente y la Luna da una vuelta alrededor de la Tierra cada 28 días poco más o menos. De modo que el movimiento de la Luna combina estos dos diferentes períodos. Por supuesto, la idea general del problema de los tres cuerpos es que esta descripción no es totalmente exacta, pero soluciones «cuasiperiódicas» de este tipo son muy habituales en problemas de muchos cuerpos. La sección de Poincaré detecta soluciones cuasiperiódicas: cuando vuelven a la superficie no inciden exactamente en el mismo punto sino que inciden en un punto que se mueve en una curva cerrada *sobre la superficie*, en pequeños pasos.

Poincaré se dio cuenta de que si todas las soluciones eran así, él sería capaz de construir series apropiadas para modelarlas cuantitativamente. Pero cuando analizó la topología de la primera aplicación de retorno, advirtió que podía ser más complicada. Dos curvas particulares, relacionadas por la dinámica, podrían cruzarse. Esto no era demasiado malo en sí mismo, pero cuando se prolongaban las curvas hasta que incidían de nuevo en la superficie, las curvas resultantes seguían teniendo que cruzarse... pero en un lugar diferente. Prolonguémoslas de nuevo, y se vuelven a cruzar. No solo eso: estas nuevas curvas que aparecen al prolongar las originales no eran realmente nuevas. Eran partes de las curvas originales. Ordenar la topología necesitó alguna lúcida reflexión, porque nadie había tomado parte antes en un juego de este tipo. Lo que emerge es una imagen muy compleja, como una red loca, en la que curvas zigzaguean repetidamente, se cruzan, y los propios zigzags zigzaguean, y así de modo sucesivo en cualquier nivel de complejidad. De hecho, el propio Poincaré se declaró sorprendido:

> *Cuando uno trata de representar la figura formada por estas dos curvas y su infinidad de intersecciones, cada una de las cuales corresponde a una solución doblemente asintótica, estas intersecciones forman una especie de red, madeja o malla infinitamente fina... Uno queda perplejo ante la complejidad de esta figura que ni siquiera intentaré dibujar.*

Ahora llamamos a esta imagen una maraña («autoconexa») homoclina (véase Figura 31). Gracias a nuevas ideas topológicas introducidas en los años sesenta del siglo pasado por Stephen Smale, ahora reconocemos en esta estructura a una vieja amiga. Su implicación más importante es que la dinámica es caótica. Aunque las ecuaciones no tienen un elemento de aleatoriedad explícito, sus soluciones son muy complicadas e irregulares, compartiendo ciertas propiedades de procesos genuinamente aleatorios. Por ejemplo, existen órbitas —la mayoría de ellas, de hecho— para las que el movimiento imita con total exactitud el repetido lanzamiento aleatorio de una moneda. El descubrimiento de que un sistema determinista —un sistema cuyo futuro entero está de modo unívoco determinado por su estado presente— puede sin embargo tener propiedades aleatorias es extraordinario y ha cambiado muchas áreas de la ciencia. Ya no suponemos automáticamente que reglas simples dan lugar a un comportamiento simple. Esto es lo que en lenguaje coloquial se conoce como teoría del caos, y todo se remonta a Poincaré y su premio Oscar.

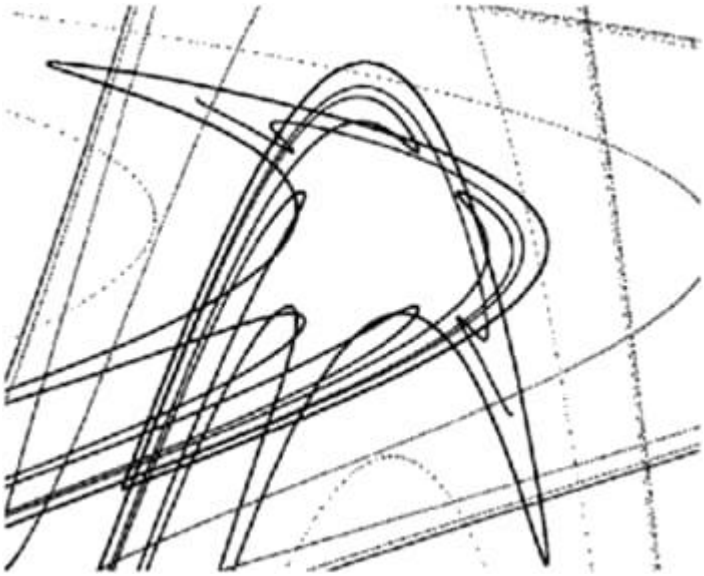


Figura 31. Parte de una maraña homoclina. Una imagen completa sería infinitamente complicada. © http://random.mostlymaths.net.

Bueno, casi todo. Durante muchos años así contaban la historia los historiadores de las matemáticas. Pero alrededor de 1990 June Barrow-Green encontró una copia impresa de la memoria de Poincaré en las profundidades del Instituto Mittag-Leffler en Estocolmo, la hojeó y se dio cuenta de que era diferente de la versión que podía encontrarse en innumerables bibliotecas matemáticas en todo el mundo. Era, de hecho, la impresión oficial de la memoria de Poincaré ganadora del premio, y había en ella un notable error. Cuando Poincaré presentó su trabajo para el premio había pasado por alto las soluciones caóticas. Él detectó el error antes de que la memoria fuera publicada, desarrolló lo que debería haber deducido —a saber, caos— y pagó (más de lo que le había reportado el premio) para que se destruyera la versión original y se imprimiera una versión corregida. Por alguna razón los archivos del

Instituto Mittag-Leffler conservaron una copia de la defectuosa versión original, pero esta quedó olvidada hasta que Barrow-Green la desenterró al publicar su descubrimiento en 1994.

Al parecer Poincaré pensaba que estas soluciones caóticas eran incompatibles con desarrollos en serie, pero eso también resultó ser falso. Era una hipótesis fácil de hacer: las series parecen demasiado regulares para representar caos; solo la topología puede hacerlo. El caos es comportamiento complicado causado por reglas simples, de modo que la inferencia no es irrefutable, pero la estructura del problema de los tres cuerpos impide definitivamente soluciones simples del tipo que derivó Newton para dos cuerpos. El problema de dos cuerpos es «integrable», lo que significa que las ecuaciones tienen suficientes cantidades conservadas, tales como energía, momento y momento angular, para determinar las órbitas. «Conservadas» significa que estas cantidades no cambian cuando los cuerpos siguen sus órbitas. Pero es sabido que el problema de los tres cuerpos no es integrable.

Pese a todo, existen soluciones en serie, pero no son universalmente válidas. Fallan para estados iniciales con momento angular cero —el momento angular es una medida de la rotación total— que son infinitamente raros porque cero es un único número entre la infinidad de los números reales. Además, no son series en la variable temporal como tal: son series en su raíz cúbica. El matemático finlandés Kart Fritiof Sundman descubrió todo esto en 1912. Algo similar es válido para el problema de *n* cuerpos, una vez más con raras excepciones, un resultado obtenido en 1991 por

Qiudong Wang. Pero para cuatro o más cuerpos no tenemos una clasificación de las circunstancias precisas en las que las series dejan de converger. Sabemos que tal clasificación debe ser muy complicada, porque existen soluciones en las que todos los cuerpos escapan al infinito, u oscilan con infinita rapidez, al cabo de un tiempo finito (véase capítulo 12). Desde el punto de vista físico estas soluciones son artificios de la hipótesis de que los cuerpos son simples puntos (masivos). Desde el punto de vista matemático, nos dicen dónde buscar comportamiento incontrolado.

Se han hecho progresos espectaculares en el problema de *n* cuerpos cuando todos los cuerpos tienen la misma masa. Esto difícilmente es una hipótesis realista en mecánica celeste, pero es razonable para algunos modelos no cuánticos de partículas elementales. El interés principal es matemático. En 1993 Christopher Moore encontró una solución del problema de los tres cuerpos en la que los tres cuerpos juegan al juego de seguir-a-mi-líder a lo largo de la misma órbita. Más sorprendente todavía es la forma de la órbita: una forma de ocho, mostrada en la Figura 32. Aunque esta órbita se cruza a sí misma, los cuerpos nunca colisionan.

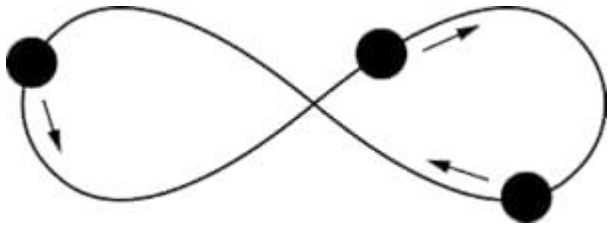


Figura 32. La coreografía en figura de ocho.

El cálculo de Moore era numérico, en un ordenador. Su solución fue redescubierta de manera independiente en 2001 por Alain Chenciner y Richard Montgomery, quienes combinaban un principio tradicional de la mecánica clásica, conocido como «mínima acción», con una topología realmente sofisticada para dar una demostración rigurosa de que tal solución existe. Las órbitas son periódicas en el tiempo: al cabo de un intervalo de tiempo dado todos los cuerpos vuelven a sus posiciones y velocidades iniciales, y después se repiten los mismos movimientos indefinidamente. Para una masa común dada hay al menos una solución de este tipo para cualquier período.

En 2000 Carles Simó realizó un análisis numérico que indicaba que la figura de ocho es estable, excepto quizá por una lenta deriva a muy largo plazo conocida como difusión de Arnold, relacionada con la geometría detallada de la aplicación de retorno de Poincaré. Para este tipo de estabilidad *casi* todas las perturbaciones llevan a una órbita muy próxima a la concernida, y conforme la perturbación se hace más pequeña, la proporción de tales perturbaciones se acerca al cien por cien. Para la pequeña proporción de perturbaciones que no se comportan de esta manera estable, la órbita se aparta muy lentamente de su localización original. El resultado de Simó fue una sorpresa, porque las órbitas estables son raras en el problema de los tres cuerpos con masas iguales. Los cálculos numéricos muestran que la estabilidad persiste incluso cuando las tres masas son ligeramente diferentes. De modo que es posible que en algún lugar en el universo tres estrellas con masas casi idénticas se estén

persiguiendo unas a otras en una figura de ocho. En 2000 Douglas Heggie estimó que el número de tales estrellas triples se encuentra entre una por galaxia y una por universo.

La figura de ocho tiene una simetría interesante. Empecemos con tres cuerpos A, B y C. Sigámoslos durante un tercio del período orbital. Entonces encontraremos tres cuerpos con las mismas posiciones y velocidades que tenían al principio, aunque ahora los cuerpos correspondientes son B, C y A. Al cabo de dos tercios del período lo mismo sucede para C, A y B. Un período completo restaura las etiquetas originales de los cuerpos. Una solución de este tipo se conoce como una coreografía: una danza planetaria en la que todos intercambian posiciones cada cierto tiempo. La evidencia numérica revela la existencia de coreografías para más de tres cuerpos: la Figura 33 muestra algunos ejemplos. Simó en particular ha encontrado un número enorme de coreografías[58].

Incluso aquí, muchas preguntas siguen sin respuesta. Carecemos de demostraciones rigurosas de la existencia de tales coreografías. Para más de tres cuerpos todas parecen ser inestables; muy probablemente esto es correcto, pero está por demostrar. La órbita con figura de ocho para tres cuerpos de masa dada y período dado parece ser única, pero tampoco se conoce ninguna demostración, aunque en 2003 Tomasz Kapela y Piotr Zglicynski proporcionaron una demostración asistida por ordenador de que es localmente única: ninguna órbita próxima funciona. Las coreografías podrían ser otro gran problema en ciernes.

[58] Puede encontrarse una animación y más información en: http://www.scholarpedia.org/article/N-body_choreographies

Entonces, ¿es estable el Sistema Solar?

Quizá sí, quizá no.

Siguiendo la gran intuición de Poincaré, la posibilidad de caos, ahora entendemos mucho más claramente las cuestiones teóricas implicadas en establecer la estabilidad. Resultan ser sutiles y complejas; y, lo que resulta irónico, no están relacionadas de ninguna manera útil con la existencia de soluciones en forma de series. Trabajos de Jürgen Moser y Vladimir Arnold han llevado a demostraciones de que varios modelos simplificados del Sistema Solar son estables para casi todos los estados iniciales, excepto quizá por el efecto de la difusión de Arnold, que impide tipos más fuertes de estabilidad en casi todos los problemas de este tipo. En 1961 Arnold demostró que un modelo idealizado de Sistema Solar es estable en este sentido, pero solo bajo la hipótesis de que los planetas tienen masas muy pequeñas comparadas con la estrella central y las órbitas son muy próximas a círculos y están muy próximas a un plano común. Por lo que respecta a una demostración rigurosa, «muy próximo» aquí significa «que difiere en un factor menor que 10^{-43}», e incluso entonces el enunciado completo es que la probabilidad de ser inestable es cero. En un argumento de perturbación de este tipo, los resultados suelen ser válidos para discrepancias mucho mayores que cualquier cosa que pueda demostrarse rigurosamente, de modo que la inferencia es que los sistemas planetarios razonablemente próximos a este ideal son con toda probabilidad estables. Sin embargo, en nuestro Sistema Solar los números relevantes son aproximadamente 10^{-3} para las

masas y 10^{-2} para la circularidad y la inclinación. Estos números superan con mucho a 10^{-43}. De modo que la aplicabilidad del resultado de Arnold era discutible. Fue en cualquier caso alentador que *algo* pudiera decirse con certeza.

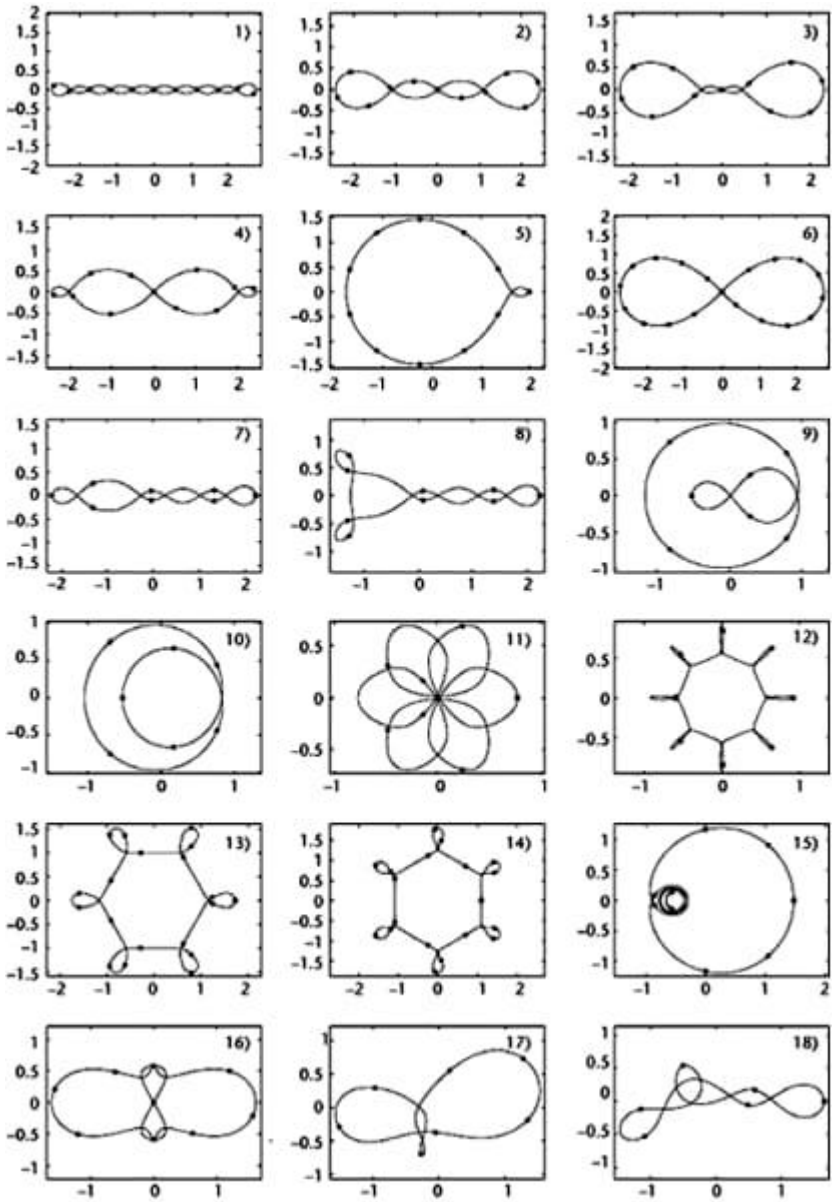


Figura 33. Ejemplos de coreografías. © Carles Simó. Del «Congreso europeo de Matemáticas», Budapest 1996, Progress in Mathematics n.º 168, Birkhäuser, Basel.

Las cuestiones prácticas en tales problemas también se han aclarado gracias al desarrollo de potentes métodos numéricos para obtener soluciones aproximadas de las ecuaciones mediante ordenador. Esta es una materia delicada porque el caos tiene una consecuencia importante: errores pequeños pueden crecer muy rápidamente y arruinar las respuestas. Nuestra comprensión teórica del caos, y de ecuaciones como las del Sistema Solar donde no hay fricción, ha llevado al desarrollo de métodos numéricos que son inmunes a muchas de las propiedades más molestas del caos. Se denominan integradores simplécticos. Utilizándolos, resulta que la órbita de Plutón es caótica. Sin embargo, eso no implica que Plutón se precipité contra el Sistema Solar provocando una catástrofe. Significa que durante un período de doscientos millones de años Plutón seguirá estando en algún lugar próximo a su órbita actual, pero no tenemos ninguna pista de en qué parte de dicha órbita estará.

En 1982 el Proyecto Lonstop de Archie Roy modeló los planetas exteriores (de Júpiter hacia fuera) en un superordenador y no encontró inestabilidad a gran escala, aunque algunos de los planetas ganaban energía a expensas de otros de maneras extrañas. Desde entonces dos grupos de investigación, en particular, han desarrollado estos métodos computacionales y los han aplicado a muchos problemas diferentes acerca de nuestro Sistema Solar. Están dirigidos por Jack Wisdom y Jacques Laskar. En 1984 el grupo de Wisdom predijo que Hiperión, un satélite de Saturno,

debería bambolearse caóticamente en lugar de girar regularmente, y observaciones posteriores lo confirmaron. En 1988, en colaboración con Gerry Sussman, el grupo construyó su propio ordenador, hecho a la medida para las ecuaciones de la mecánica celeste: el *orrery*, un planetario digital. Un *orrery* es un aparato mecánico con bielas y engranajes que simula el movimiento de los planetas, que aquí son pequeñas bolas metálicas sobre varillas[59]. La computación original siguió los próximos 845 millones de años del Sistema Solar y reveló la naturaleza caótica de Plutón. Con las siguientes, el grupo de Wisdom ha explorado la dinámica del Sistema Solar durante los siguientes miles de millones de años.

El grupo de Laskar publicó sus primeros resultados sobre el comportamiento a largo plazo del Sistema Solar en 1989, utilizando una forma promediada de las ecuaciones que se remonta a Lagrange. Aquí algo del detalle fino se difumina y se ignora. Los cálculos del grupo demostraron que la posición de la Tierra en su órbita es caótica, muy similar a lo que sucede con Plutón: si medimos dónde está la Tierra hoy y nos equivocamos en quince metros, entonces su posición en la órbita dentro de cien millones de años no puede predecirse con ninguna certeza.

Una manera de mitigar los efectos del caos es realizar muchas simulaciones, con datos iniciales ligeramente diferentes, y obtener una imagen del abanico de futuros posibles y cuan probable es cada uno de ellos. En 2009 Laskar y Mickaël Gastineau aplicaron esta técnica al Sistema Solar, con dos mil quinientos escenarios

[59] Debe su nombre al Barón de Orrey, a quien se le regaló uno en 1704.

diferentes. Las diferencias son extraordinariamente pequeñas: mover Mercurio 1 metro, por ejemplo. En aproximadamente un 1 por 100 de estos futuros, Mercurio se hace inestable: colisiona con Venus, se precipita contra el Sol, o sale despedido al espacio.

En 1999 Norman Murray y Matthew Holman investigaron la inconsistencia entre resultados como los de Arnold, que indican estabilidad, y simulaciones, que indican inestabilidad. «¿Son incorrectos los resultados numéricos, o simplemente los cálculos clásicos son inaplicables?», se preguntaron. Utilizando métodos analíticos, no numéricos, demostraron que los cálculos clásicos no son aplicables. Las perturbaciones necesarias para simular la realidad son demasiado grandes. La fuente principal de caos en el Sistema Solar es una cuasiresonancia entre Júpiter, Saturno y Urano, además de una menos importante que implica a Saturno, Urano y Neptuno. También utilizaron métodos numéricos para comprobar esta propuesta, que demostraron que el horizonte de predicción —una medida del tiempo que tardan los pequeños errores en hacerse suficientemente grandes para tener un efecto importante— es de unos diez millones de años[60]. Sus simulaciones muestran que Urano experimenta ocasionales encuentros cercanos con Saturno, cuando la excentricidad de su órbita cambia de forma caótica, y hay una posibilidad de que eventualmente sea expulsado por completo del Sistema Solar. Sin embargo, el tiempo probable es de unos 10^{18} años. El Sol se expandirá como una gigante roja mucho antes, unos cinco mil millones de años a partir de ahora, y

[60] Más formalmente, esto se denomina tiempo de Liapunov.

esto afectará a todos los planetas porque el Sol perderá el 30 por 100 de su masa. La Tierra se moverá hacia fuera, y podría escapar de ser engullida por el Sol enormemente expandido. Sin embargo, ahora se piensa que las interacciones de marea empujarán hipotéticamente a la Tierra hacia el Sol. Los océanos de la Tierra habrán hervido mucho antes. Pero puesto que la vida media típica de una especie, en términos evolutivos, no es más de unos cinco millones de años, en realidad no necesitamos preocuparnos por ninguna de estas catástrofes potenciales. Antes nos ocurrirá alguna otra cosa.

Los mismos métodos pueden utilizarse para investigar el pasado del Sistema Solar: utilizar las mismas ecuaciones y correr el tiempo hacia atrás, un simple truco matemático. Hasta hace poco los astrónomos tendían a suponer que los planetas siempre han estado próximos a sus órbitas actuales, desde que se condensaron a partir de un nube de gas y polvo alrededor del Sol naciente. De hecho, sus órbitas y composición habían sido utilizadas para inferir el tamaño y composición de dicha nube de polvo primordial. Ahora parece que los planetas no empezaron en sus órbitas actuales. Cuando la nube de polvo se rompió en grumos bajo sus propias fuerzas gravitatorias, Júpiter —el planeta más masivo— empezó a organizar las posiciones de los otros cuerpos, y estos a su vez se influyeron mutuamente. Esta posibilidad fue propuesta en 1984 por Julio Fernández y Wing-Huen Ip, pero durante un tiempo su trabajo fue visto como una curiosidad menor. En 1993 Renu Malhotra empezó a considerar seriamente cómo podían influir los cambios en la

órbita de Neptuno en los otros planetas gigantes, otros asumieron la historia y emergió una imagen de un primitivo Sistema Solar muy dinámico.

A medida que los planetas seguían agregándose, llegó un tiempo en el que Júpiter, Saturno, Urano y Neptuno estaban casi completos, pero entre ellos circulaban números enormes de planetesimales rocosos y helados, pequeños cuerpos de unos diez kilómetros de tamaño. De entonces en adelante, el Sistema Solar evolucionó mediante la migración y colisión de planetesimales. Muchos fueron expulsados, lo que redujo la energía y el momento angular de los cuatro planetas gigantes. Puesto que estos mundos tenían masas diferentes y estaban a dispares distancias del Sol, reaccionaron de formas distintas. Neptuno fue uno de los ganadores en las apuestas de energía orbital y migró hacia fuera. También lo hicieron, en menor grado, Urano y Saturno. Júpiter fue el gran perdedor y se movió hacia dentro. Pero era tan masivo que no se movió muy lejos.

Los otros cuerpos, más pequeños, del Sistema Solar también se vieron afectados por estos cambios. El esquema actual, aparentemente estable, de nuestro Sistema Solar surgió gracias a una intrincada danza de los gigantes, en la que los cuerpos más pequeños se vieron lanzados unos contra otros en un motín caótico. Entonces, ¿es estable el Sistema Solar? Probablemente no, pero no será fácil averiguarlo.

Capítulo 9

Pautas en los primos

La hipótesis de Riemann

En el capítulo 2 vimos las propiedades de los números primos como individuos, y yo las comparé con el comportamiento frecuentemente errático e impredecible de los seres humanos. Los seres humanos tienen libre albedrío; pueden tomar sus propias decisiones por sus propias razones. Los primos tienen que hacer lo que les imponga la lógica de la aritmética, pero a menudo parecen tener también una voluntad propia. Su comportamiento está gobernado por extrañas coincidencias y con frecuencia carecen de cualquier estructura razonable.

De todas formas, el mundo de los primos no se rige por la anarquía. En 1835 Adolphe Quetelet sorprendió a sus contemporáneos al encontrar genuinas regularidades matemáticas en eventos sociales que dependían de decisiones humanas conscientes o de la intervención del destino: nacimientos, matrimonios, muertes, suicidios. Las pautas eran estadísticas: no se referían a individuos, sino al comportamiento promedio de un gran número de personas. Así es como los estadísticos extraen orden a partir del libre albedrío de los individuos. Casi al mismo tiempo, los matemáticos empezaron a darse cuenta de que el mismo truco funciona con los primos. Aunque cada uno de ellos es un rudo individualista, colectivamente se atienen al imperio de la ley. Existen pautas ocultas.

Las pautas estadísticas aparecen cuando pensamos en primos comprendidos dentro de grandes intervalos. Por ejemplo, ¿cuántos primos hay hasta algún límite especificado? Esta es una pregunta muy difícil de responder exactamente, pero hay excelentes aproximaciones, y cuanto mayor es el límite, mejores se hacen estas aproximaciones. A veces la diferencia entre la aproximación y la respuesta exacta puede hacerse muy pequeña, pero habitualmente eso es demasiado pedir. La mayoría de las aproximaciones en esta área son asintóticas, lo que significa que la razón de la aproximación a la respuesta exacta puede hacerse muy próxima a 1. El error absoluto en la aproximación puede alcanzar cualquier tamaño, incluso si el error porcentual se contrae hacia cero.

Si usted se está preguntando cómo es esto posible, suponga que para alguna abstrusa propiedad de los primos la secuencia aproximada de números es la constituida por las potencias de 10:

100 10.000 1.000.000 100.000.000

pero los números exactos son:

101 10.010 1.000.100 100.001.000

donde el 1 extra se mueve un lugar hacia la izquierda en cada etapa. Entonces las razones de los números correspondientes se hacen cada vez más próximas a 1, pero las diferencias son:

1 10 100 1000

que se hacen tan grandes como queramos. Este tipo de comportamiento se da si los errores —las diferencias entre la aproximación y la respuesta exacta— crecen sin límite pero aumentan más lentamente que los propios números.

La búsqueda de fórmulas asintóticas relacionadas con los primos inspiró nuevos métodos en teoría de números, basados no en los números naturales sino en el análisis complejo. El análisis es la formulación rigurosa del cálculo infinitesimal, que tiene dos aspectos claves. Uno, el cálculo diferencial, trata del ritmo al que una cantidad, llamada una función, cambia con respecto a otra cantidad. Por ejemplo, la posición de un cuerpo depende —es una función— del tiempo, y el ritmo al que dicha posición cambia a medida que pasa el tiempo es la velocidad instantánea del cuerpo. El otro aspecto, el cálculo integral, trata de calcular áreas, volúmenes y similares sumando muchísimas piezas muy pequeñas, un proceso llamado integración. Es notable que la integración resulta ser la inversa de la diferenciación. La formulación original del cálculo infinitesimal por parte de Newton y Gottfried Leibniz requería algunas maniobras con cantidades infinitamente pequeñas, lo que planteaba cuestiones sobre la validez lógica de la teoría. Con el tiempo estas cuestiones conceptuales se resolvieron definiendo la noción de límite, un valor al que una función puede aproximarse tanto como sea requerido pero que no es necesario que se alcance realmente. Cuando se presenta de esta forma más rigurosa, la disciplina se denomina análisis.

En la época de Newton y Leibniz las cantidades de interés eran números reales, y la disciplina que nació era el análisis real. Cuando los números complejos llegaron a ser ampliamente aceptados entre los matemáticos, fue natural extender el análisis a cantidades complejas. Esta disciplina es el análisis complejo. Resultó ser extraordinariamente bello y potente. En lo que respecta al análisis, las funciones complejas se comportan mucho mejor que las reales. Tienen sus peculiaridades, desde luego, pero las ventajas de trabajar con funciones complejas compensan con creces sus desventajas.

Fue una gran sorpresa cuando los matemáticos descubrieron que las propiedades aritméticas de los números naturales podían reformularse con provecho en términos de funciones complejas. Previamente, estos dos sistemas de números habían planteado preguntas muy diferentes y utilizado métodos muy diferentes. Pero ahora, el análisis complejo, un cuerpo de técnicas en extremo potentes, podía utilizarse para descubrir propiedades especiales de funciones en teoría de números; de estas podían extraerse fórmulas asintóticas y muchas otras cosas.

En 1859 un matemático alemán, Bernhard Riemann, retomó una vieja idea de Euler y la desarrolló de una forma nueva y espectacular, al definir la denominada función zeta. Una de las consecuencias era una fórmula *exacta* para el número de primos hasta un límite dado. Era una suma infinita, pero los analistas estaban acostumbrados a esto. No era tan solo un truco inteligente pero inútil; proporcionaba nuevas y genuinas ideas sobre los

primos. Solo tenía una pega. Aunque Riemann pudo demostrar que su fórmula era exacta, sus consecuencias potenciales más importantes dependían de una simple proposición sobre la función zeta, y Riemann no pudo demostrar dicha proposición. Siglo y medio más tarde, seguimos sin poder hacerlo. Se denomina la hipótesis de Riemann, y es el Santo Grial de las matemáticas puras.

En el capítulo 2 vimos que los primos tienden a hacerse más escasos a medida que se hacen más grandes. Puesto que obtener resultados exactos sobre su distribución parecía imposible, ¿por qué no buscar en su lugar pautas estadísticas? En 1797-1798 Legendre contó cuántos primos hay hasta varios límites, utilizando tablas de primos que recientemente habían proporcionado Jurij Vega y Anton Felkel. Parece que a Vega le gustaban los cálculos largos; construyó tablas de logaritmos y en 1789 tenía el récord mundial de cálculo de π, hasta 140 cifras decimales (126 correctas). A Felkel solo le gustaba calcular primos. Su trabajo más importante es *Tafel aller Einfachen Factoren der durch 2, 3, 5 nicht theilbaren Zahlen von 1 bis 10 000 000* («Tabla de todos los factores primos de números hasta diez millones, excepto los divisibles por 2, 3 o 5») de 1776. Hay maneras fáciles de comprobar factores 2, 3 y 5, mencionadas en el capítulo 2, y por ello ahorró mucho espacio omitiendo dichos números. Legendre descubrió una aproximación empírica para el número de primos menores que un número dado x, que se denota por $\pi(x)$. Si usted solo ha visto π como un símbolo para el número 3,14159, le costará un poco acostumbrarse, pero no es difícil entender lo que se pretende, incluso si no se advierte que utilizan

tipos de letra diferentes. El texto de Legendre de 1808 sobre teoría de números afirmaba que π(*x*) parece estar muy próximo a *x*/(log *x* - 1,08366).

En una carta de 1849 al astrónomo Johann Encke, Gauss decía que cuando tenía quince años escribió una nota en sus tablas de logaritmos donde afirmaba que el número de primos menores o iguales que *x* es *x*/log *x* para *x* grande. Como sucede con muchos de sus descubrimientos, Gauss no publicó esta aproximación, quizá porque no tenía ninguna demostración de la misma. En 1838 Dirichlet informó a Gauss de una aproximación similar que él había descubierto, que se reduce a la función integral logarítmica[61]:

$$\mathrm{Li}(x) = \int_0^x \frac{dt}{\log t}$$

La razón de Li(*x*) a *x*/log *x* tiende a 1 cuando *x* se hace grande, lo que implica que si una es asintótica a π(*x*) también lo es la otra, pero la Figura 34 sugiere (correctamente) que Li(*x*) es una mejor aproximación que *x*/log *x*. La precisión de Li(*x*) es impresionante; por ejemplo:

π (1.000.000.000) = 50.847.534

Li (1.000.000.000) = 50.849.234,9

[61] Existe una variante que integra 1/log *t* de 2 a *x*, y no de 0 a *x*. Esto evita una dificultad técnica en *t* = 0, donde log *t* no está definido. A veces se utiliza la notación Li(*x*) para esta variante, y a la función definida en el texto se le llama li(*x*).

La de $x/\log x$ es peor: en este caso es 48.254.942,4.

La fórmula de aproximación —ya sea utilizando Li(x) o $x/\log x$— llegó a conocerse como el teorema de los números primos, donde «teorema» se utilizaba en el sentido de «conjetura». La búsqueda de una demostración de que estas fórmulas eran asintóticas a π(x) se convirtió en uno de los problemas clave abiertos en la teoría de números. Muchos matemáticos lo atacaron utilizando métodos tradicionales del área, y algunos llegaron cerca; sin embargo, siempre parecía haber alguna hipótesis engañosa que esquivaba la demostración. Se necesitaban nuevos métodos. Estos llegaron de una curiosa reformulación de dos de los antiguos teoremas de Euclides sobre los primos.

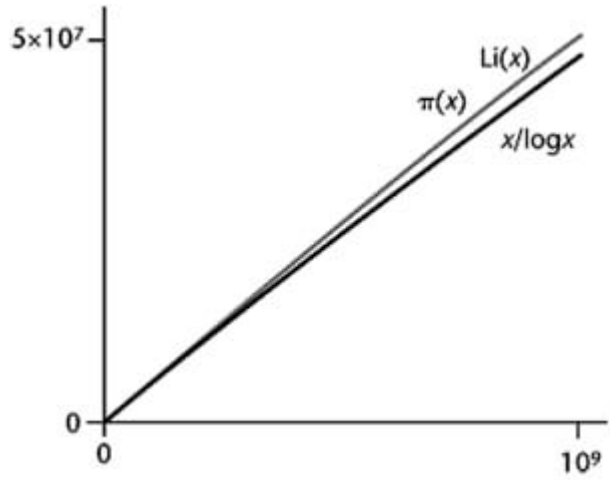


Figura 34. En esta escala π(x) y Li(x) (gris) son indistinguibles. Sin embargo, x/log x (negro) es visiblemente menor. Aquí x se representa en el eje horizontal y el valor de la función en el eje vertical.

El teorema de los números primos era una respuesta al teorema de Euclides que dice que los primos no terminan nunca. Otro teorema

euclidiano básico es la unicidad de la factorización en primos: todo entero positivo es un producto de números primos *exactamente de una manera*. En 1737 Euler se dio cuenta de que el primer teorema puede reenunciarse como una fórmula muy sorprendente en el análisis real, y el segundo enunciado se convierte en una simple consecuencia de dicha fórmula. Empezaré presentando la fórmula, y luego trataré de darle sentido. Es esta:

$$\frac{1}{1-2^{-s}} \times \frac{1}{1-3^{-s}} \times \ldots \times \frac{1}{1-p^{-s}} \times \ldots$$
$$= \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \ldots$$

Aquí p recorre todos los primos y s es constante. Euler estaba interesado sobre todo en el caso en que s es un número natural, pero su fórmula funciona también para números reales con tal de que s sea mayor que 1. Esta condición es necesaria para hacer que la serie del segundo miembro converja: que tenga un valor significativo cuando se prolonga indefinidamente.

Esta es una fórmula extraordinaria. En el primer miembro multiplicamos infinitas expresiones que solo dependen de los primos. En el segundo miembro sumamos infinitas expresiones que dependen de todos los números enteros positivos. La fórmula expresa, en lenguaje analítico, cierta relación entre números naturales y números primos. La relación más importante de este tipo es la unicidad de la factorización en primos, y esta es la que justifica la fórmula.

Voy a esbozar el paso principal para mostrar que hay una idea razonable detrás de todo esto. Utilizando el álgebra de la escuela podemos desarrollar en serie la expresión en *p*, de forma parecida al segundo miembro de la fórmula pero incluyendo solo potencias de *p*. En concreto:

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

Cuando multiplicamos todas estas series, sobre todos los primos *p*, y «desarrollamos» para obtener una suma de términos simples, obtenemos toda combinación de potencias primas; es decir, todo número natural. Cada uno aparece como el recíproco (1 dividido por) de su *s*-ésima potencia, y cada uno aparece exactamente una vez por la unicidad de la factorización en primos. Así obtenemos la serie de la derecha.

Nadie ha encontrado nunca una fórmula algebraica sencilla para esta serie, aunque hay muchas que utilizan integrales. Por eso le damos un símbolo especial, la letra griega zeta (ζ), y definimos una nueva función:

$$\zeta(s) = \frac{1}{1^s} + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{6^s} + \frac{1}{7^s} + \ldots$$

Euler no utilizó realmente el símbolo ζ, y solo consideró valores enteros positivos de *s*, pero yo llamaré a la serie anterior la función zeta de Euler. Utilizando su fórmula, Euler dedujo que existen

infinitos primos permitiendo que *s* estuviera muy próximo a 1. Si hay un número finito de primos, el primer miembro de la fórmula tiene un valor finito, pero el segundo miembro se hace infinito. Esto es una contradicción, de modo que debe haber infinitos primos. El objetivo principal de Euler era obtener fórmulas como $\zeta(2) = \pi^2/6$, que da la suma de la serie para enteros pares *s*. Él no llevó esta idea revolucionaria mucho más lejos.

Otros matemáticos detectaron lo que Euler había pasado por alto y consideraron valores de *s* que no son enteros. En dos artículos de 1848 y 1850, el matemático ruso Pafnuty Chebyshev tuvo una idea brillante: tratar de demostrar el teorema de los números primos utilizando el análisis[62]. Partió del vínculo entre los números primos y el análisis que proporcionaba la función zeta de Euler. No tuvo mucho éxito, porque supuso que *s* era real y las técnicas analíticas disponibles en el análisis real eran demasiado limitadas. Pero consiguió demostrar que cuando x es grande, la razón de $\pi(x)$ a $x/\log x$ está comprendida entre dos constantes: una ligeramente mayor que 1 y otra ligeramente menor. Tuvo una auténtica recompensa, incluso con su resultado más débil, porque le permitió demostrar el teorema de Bertrand, conjeturado en 1845: si se toma un entero y se duplica, existe un primo entre los dos.

El escenario estaba ahora preparado para Riemann. Él también reconoció que la función zeta tiene la clave para el misterio del teorema de los números primos, pero para hacer esta aproximación tenía que proponer una ambiciosa ampliación: definir la función

[62] El nombre «Pafnuty» no es habitual. Llevó a Philip Davis a escribir un libro peculiar pero absorbente: *The Thread: a Mathematical Yarn*, Harvester Press, 1983.

zeta no solo para variable real, sino también para una variable compleja. La serie de Euler es un buen lugar donde empezar. Converge para todo *s* real mayor que 1, y si se utiliza exactamente la misma fórmula para *s* complejo, entonces la serie converge siempre que la parte real de *s* sea mayor que 1. Sin embargo, Riemann descubrió que podía hacer algo mucho mejor. Utilizando un procedimiento llamado prolongación analítica extendió la definición de ζ(*s*) a *todos* los números complejos distintos de 1. Dicho valor de *s* está excluido porque la función zeta se hace infinita cuando *s* = 1.[63]

En 1859 Riemann reunió sus ideas sobre la función zeta en un artículo cuyo título se traduce como «Sobre el número de primos menores que una cantidad dada[64]». En él daba una fórmula explícita y exacta para π(x[65]). Yo voy a describir una fórmula más

---

[63] Esto se sigue de la curiosa formula de Riemann:

$$\zeta(1-s) = 2^{1-s}\pi^{-s}\operatorname{sen}\left(\frac{\pi(1-s)}{2}\right)\Gamma(s)\zeta(s)$$

donde Γ(*s*) es una función clásica llamada función gamma, definida para todo complejo *s*. El segundo miembro está definido cuando la parte real de *s* es mayor que 1.

[64] Bernhard Riemann, «Über die Anzahl der Primzahlen unter einer gegebenen Grösse», *Monatsberichte der Königlich Preußischen Akademie der Wissenchaften zu Berlin*, noviembre 1859.

[65] Riemann definió una función íntimamente relacionada.

$$\Pi(x) = \pi(x) + \frac{1}{2}\pi\left(x^{1/2}\right) + \frac{1}{3}\pi\left(x^{1/3}\right) + \frac{1}{4}\pi\left(x^{1/4}\right) + \ldots$$

que cuenta potencias primas en lugar de primos. A partir de esta podemos recuperar π(*x*). Luego demostró una fórmula exacta para esta función modificada en términos de integrales logarítmicas y una integral relacionada:

$$\Pi(x) = \operatorname{Li}(x) - \sum_{\rho}\operatorname{Li}\left(x^{\rho}\right) + \int_{x}^{\infty}\frac{dt}{t\left(t^2-1\right)\log t}$$

simple, equivalente a la de Riemann, para mostrar cómo aparecen los ceros de la función zeta. La idea es contar cuántos primos y potencias de primos hay hasta cualquier límite escogido. Sin embargo, en lugar de contar cada uno de ellos una vez, que es lo que hace π(x) en el caso de los primos, a los primos más grandes se les da un peso extra. De hecho, cualquier potencia de un primo se cuenta de acuerdo con el logaritmo de dicho primo. Por ejemplo, hasta un límite de 12 las potencias de primos son:

$$2, 3, 4 = 2^2, 5, 7, 8 = 2^3, 9 = 3^2, 11$$

de modo que la cuenta ponderada es

$$\log 2 + \log 3 + \log 2 + \log 5 + \log 7 + \log 2 + \log 3 + \log 11$$

que es aproximadamente 10,23.

Utilizando análisis, la información sobre esta forma más sofisticada de contar primos puede convertirse en información sobre la forma habitual. Sin embargo, esta forma lleva a fórmulas más sencillas, un pequeño precio que hay que pagar por la utilización del logaritmo. En estos términos, la fórmula exacta de Riemann afirma que esta cuenta ponderada hasta un límite *x* es igual a:

Aquí ∑ indica una suma sobre todos los números ρ para los que ζ(ρ) es cero, excluidos los enteros pares negativos.

$$-\sum_{\rho}\frac{x^{\rho}}{\rho}+x-\frac{1}{2}\log\left(1-x^{-2}\right)-\log 2\pi$$

donde ∑ indica una suma sobre todos los números ρ para los que ζ(ρ) es cero, excluidos los enteros pares negativos. Los primeros se denominan ceros no triviales de la función zeta. Los ceros triviales son los enteros pares negativos -2, -4, -6... La función zeta es cero en estos valores debido a la fórmula utilizada en la definición de la prolongación analítica, pero estos ceros no son importantes para la fórmula de Riemann, ni para muchas otras cosas.

Por si la fórmula asusta un poco, déjeme señalar el punto principal: una manera sofisticada de contar primos hasta un límite *x*, que puede convertirse en la manera habitual con unos pocos trucos analíticos, es *exactamente* igual a una suma extendida a todos los ceros no triviales de la función zeta de la simple expresión *xρ/ρ más una función sencilla de x*. Si usted es un analista complejo, vera inmediatamente que el teorema del número primo es equivalente a demostrar que la cuenta ponderada hasta el límite *x* es asintótica a *x*. Utilizando análisis complejo, esto será cierto si todos los ceros no triviales de la función zeta tienen partes reales entre 0 y 1. Chebyshev no pudo demostrarlo, pero llegó bastante cerca para obtener información útil.

¿Por qué son tan importantes los ceros de la función zeta? Un teorema básico en el análisis complejo establece que, sujeta a ciertas condiciones técnicas, una función de una variable compleja está completamente determinada por los valores de la variable para

los que dicha función es cero o infinito, junto con alguna otra información sobre su comportamiento en dichos puntos. Estos lugares especiales son conocidos como los ceros y los polos de la función. Este teorema no funciona en el análisis real; esta es una de las muchas razones por las que el análisis complejo se convirtió en el escenario preferido, pese a requerir la raíz cuadrada de menos uno. La función zeta tiene un polo, en $s = 1$, de modo que todo lo relativo a ella está determinado por sus ceros siempre que tengamos en mente este único polo.

Por conveniencia, Riemann trabajó sobre todo con una función relacionada, la función xi ξ(x), que está íntimamente relacionada con la función zeta y surge del método de prolongación analítica. Él comentó que

> *es muy probable que todos [los ceros de la función xi] sean reales. Pero a uno le gustaría tener una demostración rigurosa de ello; sin embargo, después de algunos intentos fugaces y fútiles, yo he dejado de lado provisionalmente la investigación de la misma, pues parece innecesaria para el próximo objetivo de mi investigación.*

Esta afirmación sobre la función xi es equivalente a otra sobre la función zeta relacionada. A saber: todos los ceros no triviales de la función zeta son números complejos de la forma $½ + it$, es decir, se encuentran en la *recta crítica* «parte real igual ½» (véase Figura 35). Esta versión de su comentario es la famosa hipótesis de Riemann.

El comentario de Riemann es más bien informal, como si la hipótesis de Riemann no fuera sumamente importante. No lo era para su programa de demostrar el teorema de los números primos. Pero para muchas otras preguntas, su importancia era crucial. De hecho, la hipótesis de Riemann es por lo general considerada como la más importante pregunta no respondida en matemáticas.

Para entender por qué, debemos seguir un poco más lejos el pensamiento de Riemann. Él tenía la vista puesta en el teorema de los números primos. Su fórmula exacta sugería un modo de conseguirlo: entender los ceros de la función zeta, o lo que es equivalente, de la función xi. No se necesita la hipótesis de Riemann plena; solo hay que demostrar que todos los ceros no triviales de la función zeta tienen partes reales entre 0 y 1. Es decir, se encuentran a menos de una distancia ½ de la recta crítica de Riemann, en la denominada banda crítica. Esta propiedad de los ceros implica que la suma sobre todos los ceros de la función zeta, en la fórmula exacta anterior, es una constante finita. Asintóticamente, para *x* grande, podría también no existir. Entre los términos de la fórmula, el único que sigue siendo importante cuando *x* se hace muy grande es *x*. Todas las cosas complicadas desaparecen asintóticamente en comparación con *x*. Por consiguiente, la cuenta ponderada es asintótica a *x*, y eso demuestra el teorema de los números primos. Así pues, irónicamente, el papel de los ceros de la función zeta es demostrar que la contribución de los ceros de la función zeta a la fórmula exacta no es importante.

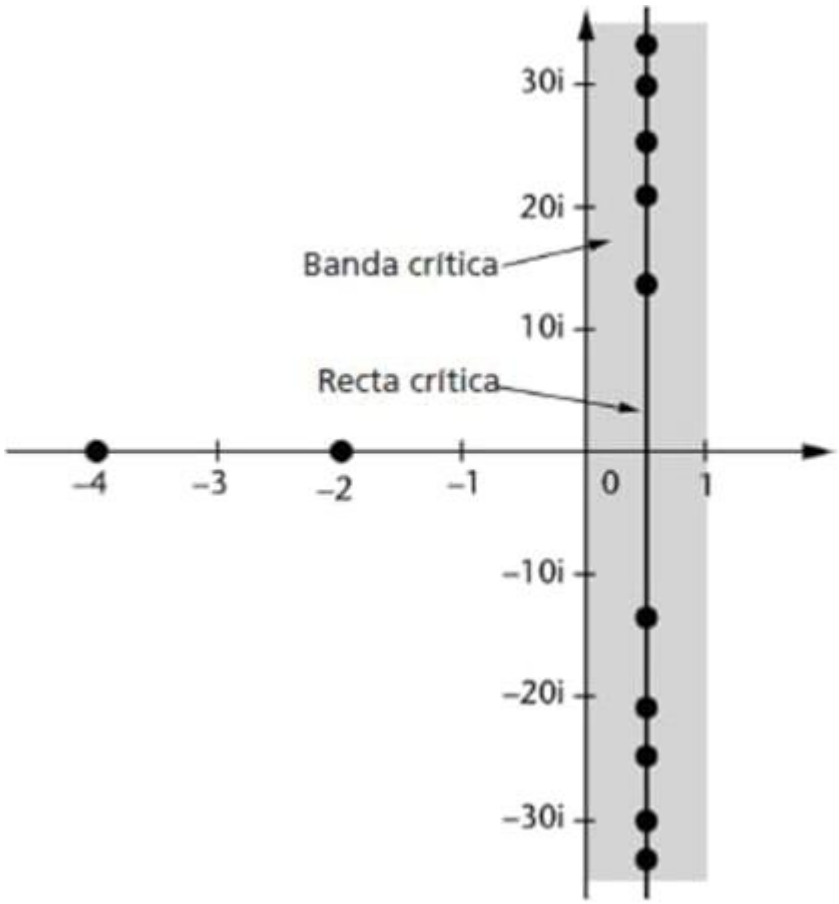


Figura 35. Ceros, recta crítica y banda crítica de la función zeta.

Riemann nunca llevó este programa hasta una conclusión. De hecho, ya no volvió a escribir sobre el tema. Pero otros dos matemáticos asumieron el reto y demostraron que la corazonada de Riemann era correcta. En 1896 Jacques Hadamard e, independientemente, Charles Jean de la Vallée Poussin dedujeron el teorema de los números primos al demostrar que todos los ceros no triviales de la función zeta se encuentran en la banda crítica. Sus demostraciones eran muy complicadas y técnicas; en cualquier caso, funcionaban. Nació una nueva y potente área de las

matemáticas: la teoría de números analítica. Tenía aplicaciones en toda la teoría de números al resolver viejos problemas y revelar nuevas pautas. Otros matemáticos encontraron demostraciones analíticas más sencillas del teorema de los números primos, y Atle Selberg y Paul Erdős descubrieron una demostración muy complicada que no requería en absoluto el análisis complejo. Pero para entonces la idea de Riemann había sido utilizada para demostrar muchos teoremas importantes, incluidas aproximaciones a muchas funciones de teoría de números. Así, su nueva demostración añadía una irónica nota a pie de página, pero por lo demás tuvo poco efecto. En 1980 Donald Newman encontró una demostración mucho más simple, que utilizaba solamente uno de los resultados más básicos del análisis complejo, conocido como teorema de Cauchy.

Aunque Riemann declaró que su hipótesis era innecesaria para sus objetivos inmediatos, resultó ser vital para muchas otras preguntas en teoría de números. Antes de discutir la hipótesis de Riemann, vale la pena echar una mirada a algunos de los teoremas que se seguirían si pudiera demostrarse que la hipótesis es cierta.

Una de las implicaciones más importantes es el tamaño del error en el teorema de los números primos. El teorema afirma que para x grande, la razón de $\pi(x)$ a Li (x) se hace cada vez más próxima a 1. Es decir, el tamaño de la diferencia entre estas dos funciones se contrae a cero, *con relación al tamaño de x*. Sin embargo, la diferencia real puede hacerse (y se hace) cada vez mayor.

Sencillamente lo hace a un ritmo más lento que la propia x[66]. Experimentos por ordenador sugieren que el tamaño del error es casi proporcional a $\sqrt{x} \log x$. Si la hipótesis de Riemann es cierta, esta afirmación puede demostrarse. En 1901 Helge von Koch demostró que la hipótesis de Riemann es lógicamente equivalente a la estimación:

$$\left|\pi(x) - \mathrm{Li}(x)\right| \leq \frac{1}{8\pi}\sqrt{x}\log x$$

para todo x ≥ 2657. Aquí las barras verticales | | indican el valor absoluto: la diferencia multiplicada por ±1 para hacerla positiva. Esta fórmula proporciona la mejor cota posible a la diferencia entre π(x) y Li(x).

La hipótesis de Riemann implica muchas estimaciones para otras funciones en teoría de números. Por ejemplo, es equivalente a que la suma de los divisores de n es menor que

$$e^{\gamma}\, n \log\log n$$

para todo $n \geq 5040$, donde $\gamma = 0{,}57721\ldots$ es la constante de Euler[67]. Estos hechos pueden parecer simples curiosidades, pero buenas estimaciones para funciones importantes son vitales para muchas

[66] Por ejemplo, $x + \sqrt{x}$ es asintótica a x: la razón es

$$(x + \sqrt{x}) \,/\, x = 1 + 1/\sqrt{x}$$

Cuando x aumenta, también lo hace $\sqrt{x}$, de modo que $1/\sqrt{x}$ tiende a 0 y la razón tiende a 1. Pero la diferencia es $\sqrt{x}$, y eso se hace cada vez mayor a medida que x aumenta. Por ejemplo, cuando x es 1 billón, $\sqrt{x}$ es 1 millón.

[67] La constante de Euler es el límite, cuando n tiende a infinito, de

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} - \log n$$

aplicaciones, y muchos teóricos de números darían su brazo derecho por demostrar cualquiera de ellas.

La hipótesis de Riemann nos dice también cuán grandes pueden ser los intervalos entre números primos consecutivos. Podemos deducir el tamaño típico de este intervalo a partir del teorema de los números primos: en promedio, el intervalo entre un primo p y el siguiente es comparable a log p. Algunos espaciados son menores, otros son mayores, y la vida de los matemáticos sería más fácil si conocieran qué tamaño podrían llegar a tener los espaciados más grandes. Harald Cramér demostró en 1936 que si la hipótesis de Riemann es cierta, el espaciado en el primo p no es mayor que una constante multiplicada por $\sqrt{p}$ log p.

La verdadera importancia de la hipótesis de Riemann es mucho más profunda. Hay generalizaciones de gran alcance, y un fuerte presentimiento de que quien pueda demostrar la hipótesis de Riemann, probablemente podrá demostrar la correspondiente hipótesis de Riemann generalizada. Lo que a su vez daría a los matemáticos mucho control sobre áreas muy amplias de la teoría de números.

La hipótesis de Riemann generalizada surge de una descripción más fina de los números primos. Todos los primos distintos de 2 son impares, y vimos en el capítulo 2 que los impares pueden clasificarse en dos tipos: los que superan en 1 a un múltiplo de 4, y los que superan en 3 a un múltiplo de 4. Se dice que son de la forma $4k + 1$ o $4k + 3$, donde k es lo que se multiplica por 4 para

obtenerlos. He aquí una corta lista de los primeros primos de cada tipo, junto con los correspondientes múltiplos de 4:

múltiplo de 4	0	4	8	12	16	20	24	28	32	36
más 1	0	5	•	13	17	•	•	29	•	37
más 3	3	7	11	•	19	23	•	•	•	•

El punto indica que el número concernido no es primo.

¿Cuántos primos de cada tipo hay? ¿Cómo están distribuidos entre los primos, o entre todos los enteros? La demostración de Euclides de que hay infinitos primos puede modificarse sin mucho esfuerzo para demostrar que existen infinitos primos de la forma $4k + 3$. Es mucho más difícil demostrar que hay infinitos primos de la forma $4k + 1$; puede hacerse, pero solamente utilizando algunos teoremas nada sencillos. La diferencia se debe a que cualquier número de la forma $4k + 3$ tiene algún factor de dicha forma; lo mismo no es siempre cierto para los números $4k + 1$.

No hay aquí nada sagrado sobre los números. Aparte de 2 y 3, todos los primos son o bien de la forma $6k + 1$ o de la forma $6k + 5$, y podemos plantear preguntas similares. Para lo que importa, todos los primos excepto 5 toman una de las formas $5k + 1$, $5k + 2$, $5k + 3$, $5k + 4$. Dejamos fuera $5k$ porque estos son múltiplos de 5, de modo que ninguno de ellos salvo 5 es primo.

No es difícil dar con una conjetura razonable para todas las cuestiones de este tipo —primos en una secuencia aritmética—. El caso $5k$ es típico. El experimento sugiere rápidamente que los

cuatro tipos enumerados tienen la misma probabilidad de ser primos. He aquí una tabla similar:

múltiplo de 5	5	10	15	20	25	30	35	40
más 1	•	11	•	•	•	31	•	41
más	7	•	17	•	•	•	37	•
más	•	13	•	23	•	•	•	43
más	•	•	19	•	29	•	•	•

De modo que debería haber infinitos primos de cada tipo individual, y en promedio alrededor de una cuarta parte de los primos, hasta algún límite dado, debería ser de cualquier forma específica.

Demostraciones sencillas muestran que algunas formas llevan a infinitos primos, demostraciones más complicadas funcionan para otras formas, pero hasta mediados del siglo XIX nadie pudo demostrar que hay infinitos primos de cada forma posible, y mucho menos que las proporciones son aproximadamente iguales. Lagrange lo supuso sin demostración en su trabajo sobre la ley de reciprocidad cuadrática —una profunda propiedad de los cuadrados respecto a un módulo primo— en 1785. Evidentemente los resultados tenían consecuencias útiles, y era el momento justo para que alguien los demostrara. En 1837 Dirichlet probó cómo adaptar las ideas de Riemann sobre el teorema de los números primos para demostrar ambas proposiciones. El primer paso consistía en definir funciones análogas a la función zeta para estos tipos de primos. Las funciones resultantes se denominan L-funciones de Dirichlet. Un ejemplo, que aparece en el caso $4k + 1/4k + 3$, es:

$$L(s,x) = 1 - 3^{-s} + 5^{-s} - 7^{-s} + 9^{-s} \ldots$$

donde los coeficientes son +1 para números de la forma $4k + 1$, -1 para números $4k + 3$, y 0 para los demás. La letra griega x se denomina un carácter de Dirichlet, y nos recuerda que utilicemos estos signos.

En el caso de la función zeta de Riemann lo que importa no es solo la serie sino también su prolongación analítica, que da a la función un significado para todos los números complejos. Lo mismo sucede con la L-función, y Dirichlet definió una prolongación analítica adecuada. Adaptando las ideas utilizadas para demostrar el teorema de los números primos, él fue capaz de demostrar un teorema análogo para primos de formas específicas. Por ejemplo, el número de primos de la forma $5k + 1$ menores o iguales que x es asintótico a $\mathrm{Li}(x)/4$, y lo mismo sucede para los otros tres casos $5k + 2$, $5k + 3$, $5k + 4$. En particular, hay infinitos primos de cada forma.

La función zeta de Riemann es un caso especial de una L-función de Dirichlet para primos de la forma $1k + 0$, es decir, todos los primos. La hipótesis de Riemann generalizada es la generalización obvia de la hipótesis de Riemann original: los ceros de cualquier L-función de Dirichlet o bien tienen parte real 1/2 o bien son «ceros triviales» con parte real o bien negativa o bien mayor que 1.

Si la hipótesis de Riemann generalizada es cierta, entonces también lo es la hipótesis de Riemann. Muchas de las consecuencias de la hipótesis de Riemann generalizada son análogas a las de la

hipótesis de Riemann. Por ejemplo, pueden demostrarse cotas de error similares para las versiones análogas del teorema de los números primos, aplicado a primos de cualquier forma específica. Sin embargo, la hipótesis de Riemann generalizada implica muchas cosas que son muy diferentes de cualquier cosa que podamos derivar utilizando la hipótesis de Riemann ordinaria. Así, en 1917 Godfrey Harold Hardy y John Edensor Littlewood demostraron que la hipótesis de Riemann generalizada implica una conjetura de Chebyshev, que afirma que (en un sentido preciso) los primos de la forma $4k + 3$ son más abundantes que los de la forma $4k + 1$. Ambos tipos son igualmente probables, a la larga, por el teorema de Dirichlet, pero eso no impide que los primos $4k + 3$ superen a los primos $4k + 1$ si se fija el juego correcto.

La hipótesis de Riemann generalizada también tiene importantes implicaciones para los testes de primalidad, tales como el test de Miller de 1976 mencionado en el capítulo 2. Si la hipótesis de Riemann generalizada es cierta, entonces el test de Miller proporciona un algoritmo eficiente. Las estimaciones de la eficiencia de tests más recientes también dependen de la hipótesis de Riemann generalizada. También hay aplicaciones importantes a la teoría de números algebraica. Recordemos del capítulo 7 que la reformulación de Dedekind de los números ideales de Kummer llevó a un concepto nuevo y fundamental, los ideales. Existe la factorización en primos en anillos de enteros algebraicos, pero quizá no sea única. La factorización en primos de ideales es mucho más rígida: existencia y unicidad son ambas válidas. Por ello tiene

sentido reinterpretar todas las preguntas sobre factores en términos de ideales. En particular, existe una noción de un «primo ideal», un análogo razonable y tratable de un número primo.

Sabiendo esto, es natural preguntar si el vínculo de Euler entre primos ordinarios y la función zeta tiene un análogo para los primos ideales. Si es así, toda la poderosa maquinaria de la teoría de números analítica se hace disponible para los números algebraicos. El caso es que puede hacerse, con implicaciones profundas y vitales. El resultado es la función zeta de Dedekind: una función semejante para cada sistema de números algebraicos. Hay un vínculo profundo entre las propiedades analíticas complejas de la función zeta de Dedekind y la aritmética de los números primos para los enteros algebraicos correspondientes. Y, por supuesto, existe un análogo de la hipótesis de Riemann: todos los ceros no triviales de la función zeta de Dedekind se encuentran en la recta crítica. Ahora la expresión «hipótesis de Riemann generalizada» incluye también esta conjetura.

Ni siquiera esta generalización es el final de la historia de la función zeta. Ha inspirado la definición de funciones análogas en otras diversas áreas de las matemáticas, desde el álgebra abstracta a la teoría de sistemas dinámicos. En todas estas áreas existen proposiciones análogas a la hipótesis de Riemann de alcance aún mayor. De algunas de ellas se ha demostrado que son ciertas. En 1974 Pierre Deligne demostró una proposición análoga semejante para variedades sobre campos finitos. Generalizaciones conocidas como funciones zeta de Selberg satisfacen una proposición análoga

a la hipótesis de Riemann. Lo mismo sucede con la función zeta de Goss. Sin embargo, existen otras generalizaciones, las funciones zeta de Epstein, para las que el análogo adecuado de la hipótesis de Riemann es falso. Aquí infinitos ceros no triviales yacen en la recta crítica, pero algunos no lo hacen, como demostró Edward Titchmarsh en 1986. Por otra parte, estas funciones zeta no tienen una fórmula de tipo producto como la de Euler, de modo que deja de parecerse a la función zeta de Riemann en lo que muy bien puede ser un aspecto crucial.

Las pruebas circunstanciales a favor de la verdad de la hipótesis de Riemann —ya sea la original o sus generalizaciones— son numerosas. Muchas cosas hermosas se seguirían de la verdad de la hipótesis. Ninguna de estas cosas ha sido refutada: hacerlo sería refutar la hipótesis de Riemann, pero no se conoce demostración ni refutación. Hay una sensación ampliamente extendida de que una demostración de la hipótesis de Riemann original también abriría el camino a una demostración de sus generalizaciones. De hecho, podría ser mejor atacar la hipótesis de Riemann generalizada en todo su esplendor, explotando la riqueza de métodos ahora disponibles, y luego deducir la hipótesis de Riemann original como un caso especial.

Hay también una enorme cantidad de pruebas experimentales a favor de la verdad de la hipótesis de Riemann, o lo que ciertamente parece una enorme cantidad hasta que alguien arroje agua fría sobre dicha afirmación. Según Carl Ludwig Siegel, Riemann calculó

los primeros pocos ceros de su función zeta numéricamente pero no publicó los resultados: están localizados en:

$$\frac{1}{2} \pm 14{,}135i \qquad \frac{1}{2} \pm 21{,}022i \qquad \frac{1}{2} \pm 25{,}011i$$

Los ceros no triviales se dan siempre en pares ± como este. He escrito aquí 1/2 en lugar de 0,5 porque la parte real se conoce *exactamente* en estos casos, explotando resultados generales en análisis complejo y propiedades conocidas de la función zeta. Lo mismo sucede con los cálculos por ordenador citados más abajo. No solo muestran que los ceros están muy cerca de la recta crítica; en realidad, están sobre ella.

En 1903 Jorgen Gram demostró numéricamente que los diez primeros (pares ± de) ceros se encuentran en la recta crítica. Para 1935 Titchmarsh había aumentado el número hasta 195. En 1936 Titchmarsh y Leslie Comrie demostraron que los primeros 1041 pares de ceros están en la recta crítica; fue la última vez que alguien hizo cálculos de este tipo a mano. Alan Turing es más conocido por su trabajo en tiempo de guerra en Bletchley Park, donde ayudó a descifrar el código alemán Enigma, y por su trabajo sobre los fundamentos de la computación y de la inteligencia artificial. Pero también se interesó en la teoría de números analítica. En 1953 descubrió un método más eficiente para calcular ceros de la función zeta, y utilizó un ordenador para deducir que los primeros 1104 pares de ceros están sobre la recta crítica. Se acumularon las pruebas de que todos los ceros hasta un límite estaban sobre la

recta crítica; el récord actual, obtenido por Yannick Saouter y Patrick Demichel en 2004, es diez billones (10^{13}). Varios matemáticos y científicos de la computación han comprobado también otros rangos de ceros. Hasta la fecha, todo cero no trivial que ha sido computado se encuentra sobre la recta crítica.

Esto podría parecer concluyente, pero los matemáticos son ambivalentes sobre este tipo de evidencias, por buenas razones. Números como diez billones pueden sonar grandes, pero en teoría de números lo que suele importar es el logaritmo del número, que es proporcional al número de dígitos. El logaritmo de diez billones está por debajo de 30. De hecho, muchos problemas dependen del logaritmo del logaritmo, o incluso el logaritmo del logaritmo del logaritmo. En dichos términos, diez billones es *minúsculo*, de modo que la evidencia numérica hasta diez billones apenas tiene peso.

Hay también alguna evidencia analítica general, que no está sujeta a esta objeción. Hardy y Littlewood demostraron que infinitos ceros se encuentran sobre la recta crítica. Selberg demostró que una proporción no nula de ceros se hallan sobre la recta crítica. Norman Levinson demostró que esta proporción es al menos un tercio, una cifra ahora mejorada hasta al menos el 40 por 100. Todos estos resultados sugieren que si la hipótesis de Riemann es falsa, los ceros que no se encuentran sobre la recta crítica son muy grandes, y muy raros. Por desgracia, la implicación más importante es que si tales excepciones existen, hallarlas será extraordinariamente difícil.

¿Por qué molestarse? ¿No debería esta evidencia numérica satisfacer a cualquier persona razonable? Por desgracia, no. No

satisface a los matemáticos, y en este caso no se trata solo de pedantería: realmente están actuando como personas razonables. En matemáticas en general, y sobre todo en teoría de números, la evidencia «experimental» aparentemente amplia suele tener mucho menos peso que el que cabría imaginar.

Una lección la proporciona la conjetura de Pólya, enunciada en 1919 por el matemático húngaro George Pólya. Él sugirió que al menos la mitad de todos los números naturales hasta cualquier valor concreto tienen un número impar de factores primos. Aquí los factores repetidos se cuentan por separado, y empezamos en 2. Por ejemplo, hasta 20 el número de factores primos se parece a la Tabla 2, donde la columna «porcentaje» da el porcentaje de números hasta este tamaño con un número impar de factores primos.

TABLA 2. Porcentajes de números, hasta un tamaño dado, que tienen un número impar de factores primos.

Número	*Factorización*	*¿Cuántos primos?*	*Porcentaje*
2	2	1	100
3	3	1	100
4	2^2	2	66
5	5	1	75
6	2×3	2	60
7	7	1	66
8	2^3	3	71
9	3^2	2	62
10	2×5	2	55
11	11	1	60
12	$2^2 \times 3$	3	63
13	13	1	66
14	2×7	2	61
15	3×5	2	57
16	2^4	4	60
17	17	1	62
18	2×3^2	3	64
19	19	1	66
20	$2^2 \times 5$	3	65

Todos los porcentajes en la última columna son mayores que el 50 por 100, y cálculos más extensos hacen razonable conjeturar que esto es siempre cierto. En 1919, sin disponer de ordenadores, los experimentos no pudieron encontrar números que refutaran la conjetura. Pero en 1958 Brian Haselgrove utilizó la teoría de números analítica para demostrar que la conjetura es falsa para algún número —menor que $1{,}845 \times 10^{361}$—, para ser exactos. Una vez que los ordenadores entraron en escena, Sherman Lehman demostró que la conjetura es falsa para 906.180.359. En 1980 Minoru Tanaka demostró que el menor de tales ejemplos es 906.150.257. De modo que pese a que la conjetura es falsa, podría

haberse acumulado evidencia experimental a su favor para casi todos los números hasta mil millones.

Además, es bueno saber que el número 906.150.257 es inusualmente interesante.

Por supuesto, los ordenadores actuales refutarían la conjetura en pocos segundos si se programaran de modo adecuado. Pero a veces ni siquiera los ordenadores sirven de ayuda. Un ejemplo clásico es el número de Skewes, donde aparentemente enormes cantidades de pruebas numéricas sugerían en principio que una famosa conjetura debería ser cierta, pese a que de hecho es falsa. Este número gigantesco aparecía en un problema íntimamente relacionado con la hipótesis de Riemann: la aproximación de π(x) por Li(*x*). Como acabamos de ver, el teorema de los números primos afirma que la razón de estas dos cantidades tiende a 1 cuando *x* se hace grande. Los cálculos numéricos parecen indicar algo más fuerte: la razón es siempre menor que 1; es decir, π(x) es menor que Li(*x*). En 2008 las computaciones numéricas de Tadej Kotnik demostraron que esto es cierto siempre que *x* sea menor que 10^{14}. Para 2012 Douglas Stoll y Demichel habían mejorado esta cota hasta 10^{18}, una cifra obtenida independientemente por Andry Kulsha. Resultados de Tomás Oliveira e Silva sugieren que puede aumentarse hasta 10^{20}.

Esto podría sonar definitivo. Es más fuerte que los mejores resultados numéricos que tenemos a favor de la hipótesis de Riemann. Pero en 1914 Littlewood demostró que esta conjetura es falsa, y lo es de forma espectacular. Conforme *x* recorre los números reales positivos, la diferencia π(x) - Li(*x*) cambia de signo (de

negativo a positivo o al revés) *infinitamente a menudo*. En particular, π(*x*) es *mayor* que Li(*x*) para algunos valores de *x* suficientemente grandes. Sin embargo, la demostración de Littlewood no daba ninguna indicación del tamaño de dicho valor.

En 1933 su estudiante, el matemático surafricano Stanley Skewes, estimó cuán grande debe ser *x*: no mayor que 10^10^10^34, donde ^ indica «elevado a la potencia». Ese número es tan gigantesco que si todos sus dígitos se imprimieran en un libro —un libro bastante aburrido, consistente en un 1 seguido por una ristra interminable de 0— el universo no sería lo bastante grande para contenerlo, incluso si cada dígito fuera del tamaño de una partícula subatómica. Además, Skewes tuvo que suponer la verdad de la hipótesis de Riemann para que su demostración funcionase. Para 1955 había encontrado una manera de evitar la hipótesis de Riemann, pero a un precio: su estimación aumentó a 10^10^10^963. Estos números son demasiado grandes incluso para el adjetivo «astronómico», pero investigación adicional los redujo a algo que podía calificarse de cosmológico. En 1966 Lehman reemplazó los números de Skewes por 10^{1165}. Te Riele lo redujo a 7×10^{370} en 1987, y en 2000 Carter Bays y Richard Hudson lo redujeron a $1{,}39822\times10^{316}$. Kuok Fai Chow y Roger Plymen rasparon un poco más, y redujeron el número a $1{,}39801\times10^{316}$. Esto puede parecer una mejora despreciable, pero es alrededor de 2×10^{313} veces menor. Saouter y Demichel hicieron una mejora adicional hasta $1{,}3971667\times10^{316}$. Mientras, en 1941 Aurel Wintner había demostrado que una proporción pequeña pero no nula de enteros

satisfacen π(x) > Li(*x*). En 2011 Stoll y Demichel computaron los primeros doscientos mil millones de ceros de la función zeta, que controlan π(x) cuando *x* es algo hasta $10^{10.000.000.000.000}$, y encontraron evidencia de que si *x* es menor que $3{,}17\times10^{114}$ entonces π(x) es menor que Li(x^{68}). De modo que para este problema particular, la evidencia hasta al menos 10^{18}, y muy posiblemente hasta 10^{114} o más, es completamente engañosa. Los veleidosos dioses de la teoría de números están divirtiéndose a costa de los seres humanos.

Durante años se han hecho muchos intentos para demostrar o refutar la hipótesis de Riemann. La página web de Matthew Watkins, «Demostraciones propuestas de la hipótesis de Riemann», da una lista de unas cincuenta de ellas desde 2000[69]. Se han encontrado errores en muchos de estos intentos, y ninguna ha sido aceptada como correcta por expertos cualificados.

Uno de los esfuerzos más ampliamente publicitados en años recientes fue el de Louis de Branges en 2002. Hizo circular un extenso manuscrito en donde pretendía deducir la hipótesis de Riemann aplicando una rama del análisis que trataba con operadores en espacios de dimensión infinita, conocida como análisis funcional. Había razones para tomar en serio a De Branges. Previamente había hecho circular una demostración de la conjetura de Bieberbach sobre desarrollos en serie de funciones complejas. Aunque su demostración original tenía errores, al final quedó

[68] Douglas A. Stoll y Patrick Demichel, «The impact of ζ(*s*) complex zeros on π(*x*) for $x < 10^{10\text{^}13}$», *Mathematics of Computation*, n.º 276 (2011), pp. 2381-2394.
[69] http://empslocal.ex.ac.uk/people/staff/mnwatkin/zeta/RHproofs.htm.

establecido que la idea subyacente funcionaba. Sin embargo, ahora parece haber buenas razones para pensar que el método propuesto por De Branges para demostrar la hipótesis de Riemann no tiene posibilidad de conseguirlo. Algunos obstáculos aparentemente fatales han sido señalados por Brian Conrey y Xian-Jin Li[70].

Quizá la mayor esperanza de una demostración proceda de modos nuevos o radicalmente diferentes de considerar el problema. Como hemos visto de modo reiterado, los avances importantes en grandes problemas surgen a menudo cuando alguien los vincula con alguna área de las matemáticas totalmente diferente. El último teorema de Fermat es un claro ejemplo: una vez que fue reinterpretado como una cuestión sobre curvas elípticas, el progreso fue rápido.

La táctica de De Branges parece ahora cuestionable, pero su enfoque es estratégicamente válido. Tiene sus raíces en una sugerencia verbal hecha alrededor de 1912 por David Hilbert, e independientemente por George Pólya. El físico Edmund Landau preguntó a Pólya si había una razón física por la que la hipótesis de Riemann debiera ser cierta. Pólya contó en 1982 que había dado con una respuesta: los ceros de la función zeta deberían estar relacionados con los valores propios de un denominado operador autoadjunto. Estos son números característicos asociados con tipos de transformaciones especiales. En física cuántica, donde tienen aplicaciones importantes, estos números determinan los niveles energéticos del sistema en cuestión, y un sencillo teorema estándar afirma que los valores propios de este tipo de operador son siempre

[70] J. Brian Conrey y Xian-Jin Li, «A note on some positivity conditions related to zeta- and *L*-functions»: http://arxiv.org/abs/math.NT/9812166

reales. Como hemos visto, la hipótesis de Riemann puede reformularse como la afirmación de que todos los ceros de la función xi son reales. Si algún operador autoadjunto tuviera valores propios que fueran los mismos que los ceros de la función xi, la hipótesis de Riemann sería una fácil consecuencia. Pólya no publicó esta idea; él no pudo encontrar tal operador, y hasta que alguien pudiera hacerlo era un castillo en el aire. Pero en 1950 Selberg demostró su «fórmula de la traza», que relaciona la geometría de una superficie con los valores propios de un operador asociado. Esto hacía la idea algo más plausible.

En 1972 Hugh Montgomery estaba visitando el Instituto de Estudio Avanzado en Princeton. Había advertido algunas sorprendentes propiedades estadísticas de los ceros no triviales de la función zeta. Se las mencionó al físico Freeman Dyson, quien inmediatamente detectó una similitud con las propiedades estadísticas de matrices hermíticas aleatorias, otro tipo especial de operador utilizado para describir sistemas cuánticos tales como núcleos atómicos. En 1999 Alain Connes dio con una fórmula de la traza, similar a la de Selberg, cuya validez implicaría la verdad de la hipótesis de Riemann generalizada. Y en 1999 los físicos Michael Berry y Jon Keating sugirieron que el operador requerido podría surgir al cuantizar un concepto bien conocido de la física clásica relacionado con el momento lineal. La conjetura de Berry resultante puede verse como una versión más específica de la conjetura de Hilbert-Pólya.

Estas ideas, que relacionan la hipótesis de Riemann con áreas nucleares de la física matemática, son notables. Muestran que el

progreso puede venir, tal vez, de áreas de las matemáticas aparentemente no relacionadas, y despierta esperanzas de que la hipótesis de Riemann pueda resolverse algún día. Sin embargo, aún no han llevado a ningún avance trascendental y definitivo que nos anime a pensar que la solución está a la vuelta de la esquina. La hipótesis de Riemann sigue siendo uno de los enigmas más desconcertantes e irritantes en el conjunto de las matemáticas.

Hoy hay una nueva razón para tratar de demostrar la hipótesis de Riemann: un premio sustancial.

No hay un premio Nobel en matemáticas. El premio más distinguido en matemáticas es la medalla Fields, o mejor dicho la Medalla Internacional para Descubrimientos Sobresalientes en Matemáticas. Debe su nombre al matemático canadiense John Fields, que dotó el premio en su testamento. Cada cuatro años, en el Congreso Internacional de Matemáticos, hasta cuatro de los más destacados matemáticos jóvenes (menores de cuarenta años) del mundo reciben una medalla de oro y un premio en metálico, actualmente quince mil dólares. Por lo que respecta a los matemáticos, la medalla Fields es equivalente en prestigio a un premio Nobel.

Muchos matemáticos consideran que la falta de un Nobel en su disciplina es una buena cosa. En la actualidad un premio Nobel asciende a algo más de un millón de dólares, una cantidad que fácilmente podría distorsionar los objetivos de la investigación y llevar a discusiones sobre prioridad. Sin embargo, la ausencia de un premio matemático importante también puede haber distorsionado la percepción que tiene el público del valor y la utilidad de las

matemáticas. Es fácil imaginar que si nadie está dispuesto a pagar por ello, no puede ser de mucho valor.

Recientemente han nacido dos nuevas recompensas matemáticas de gran prestigio. Una es el premio Abel, concedido anualmente por la Academia Noruega de Ciencias y Letras, y así llamado en honor del gran matemático noruego Niels Henrik Abel. El otro nuevo galardón consiste en los siete premios del milenio del Instituto Clay de Matemáticas. El Instituto Clay fue fundado por Landon Clay y su esposa Lavinia. Landon Clay es un hombre de negocios norteamericano que opera con fondos de pensiones, y que tiene amor, y respeto, por las matemáticas. En 1999 creó una nueva fundación para las matemáticas en Cambridge, Massachusetts, que convoca reuniones, concede becas de investigación, organiza conferencias públicas y administra un premio de investigación anual.

En 2000 *sir* Michael Atiyah y John Tate, destacados matemáticos en Gran Bretaña y Estados Unidos, anunciaron que el Instituto Clay de Matemáticas había establecido un nuevo premio con el objetivo de animar a la solución de siete de los más importantes problemas abiertos en matemáticas. Serían conocidos como los problemas del milenio, y una solución adecuadamente publicada y evaluada de cualquiera de ellos valdría un millón de dólares. En conjunto, estos problemas dirigían la atención a algunas de las cuestiones centrales sin respuesta en matemáticas, cuidadosamente seleccionadas por algunos de los matemáticos de más prestigio del mundo. El premio sustancial transmite un mensaje muy claro al público: las

matemáticas son valiosas. Todos los implicados son conscientes de que su valor intelectual puede ser más profundo que el mero dinero, pero un premio en metálico ayuda a concentrar la mente. El problema del milenio más conocido, y uno de los que se remonta más atrás en la historia, es la hipótesis de Riemann. Es la única cuestión que aparece tanto en la lista de Hilbert de 1900 como en la lista de problemas del milenio. Los otros seis problemas del milenio se discuten en los capítulos 10-15. Los matemáticos no están especialmente obsesionados por los premios, y trabajarían en la hipótesis de Riemann aunque no hubiera uno. Una idea nueva y prometedora sería toda la motivación que necesitaran.

Vale la pena recordar que las conjeturas, por consagradas que estén, pueden no ser ciertas. Hoy día la mayoría de los matemáticos parecen pensar que finalmente se encontrará una demostración de la hipótesis de Riemann, No obstante, algunos piensan que puede ser falsa: en algún lugar en las tierras vírgenes de los números muy grandes puede esconderse un cero que no se sitúa en la recta crítica. Si existe un «contraejemplo» semejante es probable que sea muy, muy grande.

Sin embargo, las opiniones cuentan poco en las fronteras de las matemáticas. La intuición experta suele ser muy buena, pero ha habido muchas ocasiones en que resultó errónea. La sabiduría convencional puede ser convencional y sabia, sin ser verdadera. Littlewood, uno de los grandes expertos en análisis complejo, era inequívoco: en 1962 dijo que estaba seguro de que la hipótesis de Riemann era falsa, y añadía que no había ninguna razón imaginable

para que fuera cierta. ¿Quién tiene razón? Solo nos queda esperar y ver.

Capítulo 10
¿Qué forma tiene una esfera?
La conjetura de Poincaré

Henri Poincaré fue uno de los más grandes matemáticos de finales del siglo XIX, un tipo algo excéntrico pero muy sagaz. Llegó a ser miembro del Bureau des Longitudes de Francia, cuyo objetivo era mejorar la navegación, la medida del tiempo y la medida de la Tierra y los planetas. Este puesto le llevó a proponer el establecimiento de zonas horarias internacionales; también le inspiró ideas sobre la física del tiempo, anticipando algunos de los descubrimientos de Einstein en relatividad especial. Poincaré dejó su marca en todo el paisaje matemático, desde la teoría de números hasta la física matemática.

En particular, fue uno de los fundadores de la topología, las matemáticas de las transformaciones continuas. Aquí, en 1904, él tropezó con una pregunta sencilla en apariencia tras darse cuenta tardíamente de que había dado por supuesta la respuesta en un trabajo anterior, pero no pudo encontrar una demostración. «Esta pregunta nos llevaría por un camino equivocado», escribió. Más bien evitaba la cuestión real; no le estaba llevando *a ninguna parte*. Aunque él lo planteaba como una pregunta, el problema llegó a conocerse como la conjetura de Poincaré porque todo el mundo esperaba que la respuesta fuera «sí». Es otro de los siete problemas del milenio del premio Clay, y lo es con razón, porque resultó ser uno de los problemas más desconcertantes en el conjunto de la

topología. La pregunta de Poincaré fue finalmente respondida en 2002 por un joven ruso, Grigori Perelman. La solución introdujo un montón de nuevas ideas y métodos, tantos que la comunidad matemática necesitó algunos años para digerir la demostración y aceptar que era correcta.

Por su éxito Perelman fue galardonado con una medalla Fields, el premio matemático más prestigioso, pero lo rechazó. No quería publicidad. Se le ofreció el premio Clay de un millón de dólares por demostrar la conjetura de Poincaré, y renunció al mismo. Tampoco quería dinero. Lo que él quería era que su trabajo fuera aceptado por la comunidad matemática. Finalmente lo fue, pero por desgracia, y por razones comprensibles, eso tomó un tiempo. Siempre fue poco realista esperar aceptación sin publicidad o sin la oferta de premios. Pero estas consecuencias inevitables del éxito no se adaptaban al carácter en ocasiones huraño de Perelman.

Ya encontramos la topología en relación con el teorema de los cuatro colores, y entonces recurrí al cliché «geometría de la lámina elástica». La geometría de Euclides trata con líneas rectas, círculos, longitudes y ángulos. Tiene lugar en un plano, o en un espacio de tres dimensiones cuando se hace más avanzada. Un plano es como una hoja de papel infinita y comparte una propiedad del papel: no se estira, ni se contrae, ni se dobla. Se puede enrollar el papel en un tubo, y se puede estirar o contraer un poco, especialmente si le cae café encima. Pero no se puede deformar una hoja de papel y hacerla redonda como una esfera sin crear grietas. Desde el punto de vista matemático, el plano euclídeo es rígido. En la geometría de Euclides

dos objetos —triángulos, cuadrados, círculos— son iguales si se puede transformar uno de ellos en el otro mediante un movimiento rígido. Y «rígido» significa que las distancias no cambian.

¿Qué pasa si se utiliza en una lámina elástica su lugar de papel? La lámina se estira, se dobla, y con poco esfuerzo puede comprimirse. Longitudes y ángulos no tienen significado fijo en una lámina elástica. De hecho, si es suficientemente elástica, tampoco lo tienen triángulos, cuadrados o círculos. Se puede deformar un triángulo en una lámina elástica para darle una esquina extra. Incluso se le puede convertir en un círculo (véase Figura 36). Cualesquiera que sean los conceptos de la geometría de la lámina elástica, no incluyen los conceptos euclídeos tradicionales.

Figura 36. Deformación topológica de un triángulo hasta un círculo.

Podría parecer que la geometría en una lámina elástica sería tan flexible que nada tendría un significado fijo, en cuyo caso pocas cosas sustanciales podrían demostrarse. No es así. Dibujemos un triángulo y situemos un punto en su interior. Si se estira y deforma la lámina hasta que el triángulo se convierte en un círculo, hay una propiedad del diagrama que no cambia: el punto sigue estando dentro. Cierto es que ahora está dentro de un círculo, y no de un triángulo, pero no está *fuera*. Para desplazar el punto al exterior

tendríamos que rasgar la lámina, lo que rompe las reglas de este juego concreto.

Hay otra propiedad que también sobrevive a la distorsión. Un triángulo es una curva cerrada simple. Es una línea que se une a sí misma de modo que no hay extremos libres y no se cruza a sí misma. Una figura de ocho es una curva cerrada pero no es simple, pues se cruza a sí misma. Cuando se deforma la lámina elástica el triángulo puede cambiar de forma, pero siempre seguirá siendo una curva cerrada. No hay manera de convertirlo en una figura de ocho, por ejemplo, sin rasgar la lámina.

En topología tridimensional, la totalidad del espacio se hace elástica. No como un bloque de goma, que recupera su forma original si se deja suelto, sino como un gel que puede deformarse sin ofrecer ninguna resistencia. Un espacio topológico es infinitamente deformable; se puede tomar una región del tamaño de un grano de arroz y ampliarla hasta el tamaño del Sol. Se pueden estirar tentáculos hasta que la región tenga la forma de un pulpo. Lo único que no está permitido hacer es introducir algún tipo de discontinuidad. No se debe rasgar el espacio, ni realizar ningún tipo de distorsión que aísle puntos vecinos.

¿Qué propiedades de una forma en el espacio sobreviven a todas las deformaciones continuas? No la longitud, ni el área, ni el volumen. Pero estar anudado sí lo hace. Si se hace un nudo en una curva y se unen los extremos para formar un lazo, entonces el nudo no puede escapar. Por mucho que deformemos el espacio, la curva sigue anudada. Así que estamos trabajando con un nuevo tipo de

geometría en la que los conceptos importantes y significativos parecen bastante difusos: «interior», «cerrado», «simple», «anudado». Esta nueva geometría tiene un nombre respetable: topología. Puede parecer más bien esotérico, quizá incluso absurdo, pero ha resultado ser una de las áreas importantes de las matemáticas del siglo XX, y sigue siendo igualmente vital en las del siglo XXI. Y una de las personas principales a quien se lo tenemos que agradecer es a Poincaré.

La historia de la topología empezó a despegar casi un siglo antes de Poincaré, en 1813. Simon Antoine Jean L'Huillier, un matemático suizo, no incendió ni mucho menos el mundo de las matemáticas durante su vida, pese a que rechazó una gran suma de dinero que un pariente había prometido pagarle si entraba en la Iglesia. L'Huillier prefirió hacer una carrera en las matemáticas. Se especializó en un remanso matemático: el teorema de Euler sobre los poliedros. En el capítulo 4 encontramos este resultado curioso y aparentemente aislado: si un poliedro tiene C caras, V vértices y A aristas, entonces $C - A + V = 2$. L'Huillier dedicó buena parte de su carrera a investigar variantes de esta fórmula, y visto en retrospectiva dio un paso vital en la dirección de la topología cuando descubrió que la fórmula de Euler es a veces falsa. Su validez depende de la forma cualitativa del poliedro.

La fórmula es correcta para poliedros sin agujeros, que pueden dibujarse sobre la superficie de una esfera o deformarse continuamente en una forma de ese tipo. Pero cuando el poliedro tiene agujeros, la fórmula falla. Por ejemplo, un portafotos hecho de

madera con una sección rectangular tiene 16 caras, 32 aristas y 16 vértices; aquí $C - A + V = 0$. L'Huillier modificó la fórmula de Euler para cubrir estos poliedros más exóticos: si hay g agujeros, entonces $C - A + V = 2 - 2g$. Este fue el primer descubrimiento de un importante invariante topológico: una cantidad asociada con un espacio que no cambia cuando el espacio es deformado de forma continua. El invariante de L'Huillier proporciona una manera rigurosa de contar cuántos agujeros tiene una superficie sin necesidad de definir «agujero». Esto es útil, porque el concepto de un agujero es delicado. Un agujero no es parte de la superficie, ni es la región fuera de la superficie. Parece ser una propiedad de cómo se sitúa la superficie en su espacio circundante. Pero el descubrimiento de L'Huillier muestra que lo que interpretamos como el *número* de agujeros es una propiedad intrínseca, independiente de cualquier espacio circundante. No es necesario definir agujeros y luego contarlos; de hecho, es mejor no hacerlo.

Después de L'Huillier, la siguiente figura clave en la prehistoria de la topología es Gauss. Él encontró otros varios invariantes topológicos cuando trabajaba en diversas áreas nucleares de las matemáticas. Su trabajo en análisis complejo, especialmente la demostración de que toda ecuación polinómica tiene al menos una solución en números complejos le llevó a considerar el índice de una curva en el plano: cuántas veces gira alrededor de un punto dado. Problemas en electricidad y magnetismo llevaron al número de enlaces de dos curvas cerradas: cuántas veces una de ellas atraviesa a la otra. Estos y otros ejemplos llevaron a Gauss a

preguntarse si podría existir alguna rama de las matemáticas aún no descubierta que proporcionara una forma sistemática de entender propiedades cualitativas de figuras geométricas. No publicó nada sobre el tema, pero lo mencionó en cartas y manuscritos.

También pasó la idea a su estudiante Johann Listing y a su ayudante August Möbius. He mencionado la banda de Möbius, una superficie con un solo lado y también un solo borde, que él publicó en 1865, y puede encontrarse en la Figura 9 del capítulo 4. Möbius señaló que «tener un solo lado», aunque intuitivamente claro, es difícil de describir de forma precisa, y propuso una propiedad relacionada que podía definirse con todo rigor. Esta propiedad era la orientabilidad. Una superficie es orientable si se puede recubrir con una red de triángulos, con flechas que circulan alrededor de cada triángulo, de modo que cuando quiera que dos triángulos tengan un lado común las flechas apunten en direcciones opuestas. Esto es lo que sucede, por ejemplo, si se dibuja una red en un plano y se hace que todas las flechas corran en sentido de las agujas del reloj. En una banda de Möbius no existe tal red.

La primera publicación de Listing en topología llegó antes, en 1847. Su título era *Vorstudien zur Topologie* («Lecciones de topología»), y fue el primer texto en emplear dicha palabra. Él había estado utilizando el término de manera informal durante aproximadamente una década. Otro término utilizado en esa época es la expresión latina *analysis situs*, «análisis de posición», pero con el tiempo perdió favor. El libro de Listing contiene pocas cosas importantes,

pero establece una noción básica: cubrir una superficie con una red de triángulos. En 1861, cuatro años después de Möbius, él describió la banda de Möbius y estudió la conectividad, esto es, si un espacio puede dividirse en dos o más partes inconexas. Desarrollando el trabajo de Listing, varios matemáticos, entre ellos Walther von Dyck, reunieron una completa clasificación topológica de superficies, suponiéndolas cerradas (sin borde) y compactas (de extensión finita). La respuesta es que toda superficie orientable es topológicamente equivalente a una esfera, a la que se le han añadido un número finito *g* de asas (véase Figura 11, izquierda y derecha, en capítulo 4). El número *g* se llama el género de la superficie, y es el que determina el invariante de L'Huillier. Si $g = 0$ tenemos la esfera, y si $g > 0$ obtenemos un toro con *g* agujeros. Una secuencia de superficies similar, que empieza con la más simple superficie no orientable, el plano proyectivo, clasifica todas las superficies no orientables. El método se extendió para permitir también superficies con bordes. Cada borde es un lazo cerrado, y la única información extra necesaria es cuántos de estos lazos se dan.

La conjetura de Poincaré tendrá más sentido si primero echamos una mirada a una de las técnicas básicas empleadas en la clasificación de superficies. Antes describí la topología en términos de deformar una forma hecha de goma o de gel, y resalté la necesidad de utilizar transformaciones *continuas*. Irónicamente, una de las técnicas centrales en topología incluye lo que a primera vista parece ser una transformación discontinua: cortar la forma en piezas. Sin embargo, la continuidad se recupera mediante una serie

de reglas que describen qué pieza se une con qué otras y de qué manera. Un ejemplo es la forma en que definimos un toro identificando lados opuestos de un cuadrado (véase Figura 12 del capítulo 4).

Identificar puntos que parecen ser distintos nos permite representar espacios topológicos complicados utilizando ingredientes más simples. Un cuadrado es un cuadrado, pero un cuadrado con reglas de identificación puede ser un toro, una botella de Klein, un cilindro, una banda de Möbius o un plano proyectivo, dependiendo de la elección de reglas (véase Figura 37). Así que cuando expliqué una transformación continua en términos de estirar y doblar una lámina elástica, pedí más de lo que en rigor es necesario. También se nos permite cortar la lámina, en una etapa intermedia, siempre que al final unamos los bordes de nuevo exactamente como estaban al principio, o bien especificamos reglas que tienen el mismo efecto. Por lo que respecta a la topología, establecer una regla para pegar bordes es lo mismo que implementar realmente la regla. Siempre que no olvidemos la regla en cualquier otra cosa que hagamos luego.

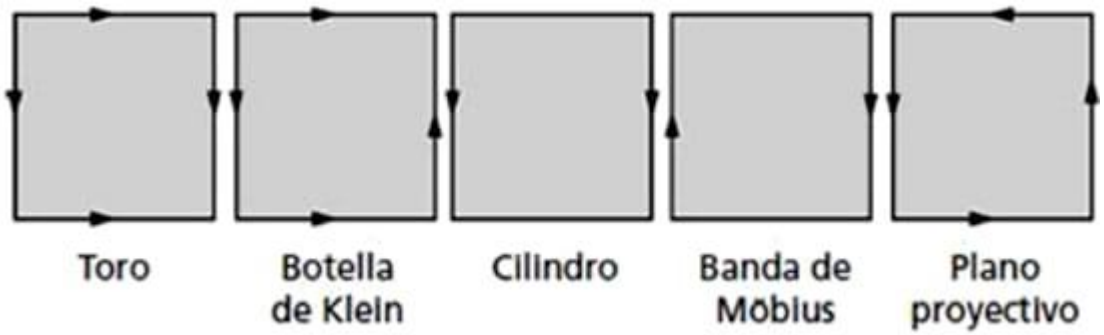


Figura 37. Cinco espacios topológicos diferentes obtenidos identificando bordes opuestos de un cuadrado de diversas maneras.

El método tradicional para clasificar superficies empieza dibujando una red de triángulos en la superficie. Luego cortamos las aristas suficientes para aplanar los triángulos y formar un polígono. Reglas de unión, derivadas de cómo hicimos los cortes, especifican cómo identificar varias aristas del polígono, lo que reconstruye la superficie original. En este punto, toda la topología interesante está implícita en las reglas de unión. La clasificación se demuestra manipulando las reglas algebraicamente y transformándolas en reglas que definen un toro de g agujeros o una de las superficies no orientables análogas. La topología moderna tiene otras maneras de conseguir el mismo resultado, pero a menudo utiliza este tipo de construcción de «corta y pega». El método se generaliza sin dificultad a espacios de cualquier dimensión, pero es demasiado restringido para llevar a una clasificación de espacios topológicos de dimensiones más altas sin ayuda adicional.

Alrededor de 1900, Poincaré estaba desarrollando el trabajo anterior sobre la topología de superficies para conseguir una técnica mucho más general que se aplicara a espacios con cualquier número de dimensiones. El motivo principal de su investigación era descubrir invariantes topológicos: números de fórmulas algebraicas asociadas con espacios, que permanecen sin cambios cuando el espacio se deforma de modo continuo. Si dos espacios tienen diferentes

invariantes, entonces uno no puede deformarse para dar el otro, de modo que son distintos topológicamente.

Él partió de la generalización que hizo en 1870 el matemático italiano Enrico Betti del invariante topológico de L'Huillier $C - A + V$ (que es ahora conocido de manera bastante injusta como la característica de Euler) a espacios de más altas dimensiones. Betti había advertido que el máximo número de curvas cerradas que pueden dibujarse en una superficie de género g, sin dividirla en piezas inconexas, es $g - 1$. Esta es otra manera de caracterizar topológicamente la superficie. Él generalizó esta idea a «números de conectividad» de cualquier dimensión, que Poincaré llamó números de Betti, un término aún en uso actualmente. El número de Betti k-dimensional cuenta el número de agujeros k-dimensionales en el espacio.

Poincaré desarrolló los números de conectividad de Betti en un invariante más sensible llamado homología, que tiene mucha más estructura algebraica. Discutiremos la homología con más detalle en el capítulo 15. Baste decir que examina colecciones de «caras» multidimensionales en este tipo de red y pregunta cuáles de ellas forman la frontera de un disco topológico. Un disco no tiene agujeros, a diferencia de un toro, de modo que podemos estar seguros de que dentro de cualquier colección de caras que constituye una frontera no hay agujeros. Recíprocamente, podemos detectar agujeros confrontando colecciones de caras que no forman fronteras con colecciones que sí lo hacen. De esta manera podemos construir una serie de invariantes de un espacio, conocidos como

sus grupos de homología. «Grupo» aquí es un término del álgebra abstracta; significa que dos objetos cualesquiera en el grupo pueden combinarse para dar alguna otra cosa en el mismo grupo, de una manera que está sometida a varias bonitas reglas algebraicas. Diré algo más adelante cuando necesitemos esta idea. Existe uno de estos grupos para cada dimensión de 0 a *n*, y para cada espacio obtenemos una serie de invariantes topológicos con todo tipo de fascinantes propiedades algebraicas.

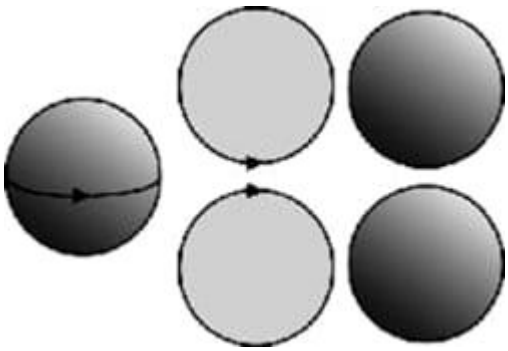
Listing había clasificado todas las superficies topológicas: espacios de dimensión 2. El siguiente caso obvio era examinar la dimensión 3. Y el espacio más simple para empezar era una esfera. En el lenguaje cotidiano la palabra «esfera» tiene dos significados diferentes: puede ser una bola redonda maciza, o solo la superficie de la bola. Cuando se trabaja en topología de superficies la palabra «esfera» se interpreta siempre en el segundo sentido: la superficie infinitamente delgada de una bola. Además, el interior de la esfera no se considera parte de ella: es solo una consecuencia de la forma usual en que insertamos una superficie esférica en el espacio. Intrínsecamente, todo lo que tenemos es una superficie, topológicamente equivalente a la superficie de una bola. Se puede considerar la esfera como una bola hueca con una piel infinitamente fina.

El análogo tridimensional «correcto» de una esfera, llamado una 3-esfera, no es una bola maciza. Una bola maciza es 3-dimensional, pero tiene una frontera: su superficie, la esfera. Una esfera no tiene una frontera, y tampoco debería tenerla su análogo 3-dimensional.

La manera más simple de definir una 3-esfera es imitar la geometría de coordenadas de una esfera ordinaria. Esto lleva a un espacio que es un poco difícil de visualizar: yo no puedo mostrarle un modelo en tres dimensiones porque la 3-esfera —incluso si solo tiene tres dimensiones— no está insertada en el espacio 3-dimensional ordinario. En su lugar, está inserta en el espacio 4-dimensional.

La esfera unidad usual, en el espacio 3-dimensional, consiste en todos los puntos que están a distancia 1 de un punto dado: el centro. Análogamente, la 3-esfera unidad en el espacio 4-dimensional consiste en todos los puntos que están a distancia unidad del centro. En coordenadas podemos escribir una fórmula para este conjunto utilizando una generalización del teorema de Pitágoras para definir la distancia[71]. De manera más general, una 3-

[71] La 3-esfera unidad comprende todos los puntos con coordenadas (x, y, z, w) tales que $x^2 + y^2 + z^2 + w^2 = 1$. Hay varias maneras de hacer la 3-esfera más intuitiva. Todas pueden entenderse por analogía con una 2-esfera y comprobarse utilizando geometría de coordenadas. Una descripción semejante («bola sólida con todos los puntos de la superficie identificados») se da en el texto, y la Figura 57 muestra otra. Para establecer la analogía, observemos que si cortamos una 2-esfera a lo largo de su ecuador obtenemos dos semiesferas. Cada una de ellas se aplana hasta dar un disco, y esto es una deformación continua. Para reconstruir la 2-esfera, simplemente identificamos puntos correspondientes en las fronteras de estos dos discos. En cierto sentido hemos hecho un mapa de la 2-esfera utilizando dos discos planos, de forma parecida a como los cartógrafos crean una proyección plana de nuestro planeta redondo. Podemos construir una 3-esfera utilizando un procedimiento análogo. Tomamos dos bolas sólidas e identificamos puntos correspondientes en sus superficies. Ahora ambos tienen la misma superficie (porque identificamos las dos superficies), y es una 2-esfera. Forma el «ecuador» de la 3-esfera.



esfera es *cualquier* espacio que es topológicamente equivalente a la 3-esfera unidad, igual que todo tipo de versiones deformadas de una 2-esfera unidad son 2-esferas topológicas, y por supuesto lo mismo sucede en dimensiones más altas.

Si usted no queda satisfecho con esto y quiere una imagen más geométrica, ensaye esta. Una 3-esfera puede representarse como una bola maciza cuya superficie entera se identifica como un solo punto. Este es otro ejemplo de una regla de unión, y en este caso es análoga a una manera de convertir un disco circular en una 2-esfera. Si rodeamos con una cuerda el borde de un disco de tela y tiramos de ella, como para cerrar una bolsa, el resultado es topológicamente igual que una 2-esfera. Realicemos ahora la operación análoga en una bola maciza, pero como es usual, no tratemos de visualizar el resultado: tan solo pensemos en una bola maciza e implementemos conceptualmente las reglas de unión.

En cualquier caso, Poincaré estaba muy interesado en la 3-esfera porque era presumiblemente el espacio topológico 3-dimensional más simple que no tenía frontera y era de extensión finita. En 1900 publicó un artículo en el que afirmaba que los grupos de homología eran un invariante lo bastante potente para caracterizar topológicamente la 3-esfera. En concreto, si un espacio topológico 3-dimensional tiene los mismos grupos de homología que una 3-esfera, entonces es topológicamente equivalente (puede ser deformado de forma continua) a una 3-esfera. Para 1904, sin

Figura 57. Cómo hacer una 3-esfera. Izquierda: Cortar una 2-esfera en semiesferas. Centro: Reconstruir la 2-esfera a partir de las dos mitades uniendo los bordes. Derecha: Por analogía, unir idealmente las superficies de dos bolas de modo que puntos correspondientes se consideran idénticos. Esto da una 3-esfera.

embargo, había descubierto que esta afirmación es falsa. Existe al menos un espacio 3-dimensional que no es una 3-esfera pero tiene los mismos grupos de homología que una 3-esfera. El espacio era un triunfo para la filosofía de las reglas de unión, y la demostración de que no era una 3-esfera incluía la creación de un nuevo invariante, necesariamente más potente que la homología.

Primero, el espacio. Se llama espacio dodecaédrico de Poincaré, porque una construcción moderna utiliza un dodecaedro macizo. Poincaré no era consciente de esta relación con un dodecaedro: él unió dos toros macizos de una manera muy oscura. La interpretación como dodecaedro fue publicada en 1933, unos veintiún años después de la muerte de Poincaré, por Herbert Seifert y Constantin Weber, y es mucho más fácil de comprender. La analogía que hay que tener en mente es la construcción de un toro uniendo lados opuestos de un cuadrado. Como siempre, no trate de *hacer* la unión; solamente recuerde que puntos correspondientes son considerados el mismo. Ahora hagamos lo mismo, pero utilizando caras opuestas de un dodecaedro (véase Figura 38).

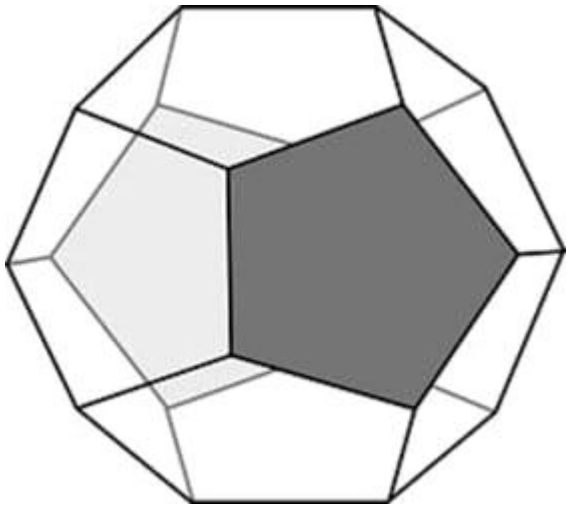


Figura 38. Para hacer el espacio dodecaédrico de Poincaré, se toma un dodecaedro y se unen todos los pares de caras opuestas (tales como el par sombreado), con un giro para que ajusten.

Los pitagóricos conocían los dodecaedros, hace 2500 años. La frontera de un dodecaedro consiste en doce pentágonos regulares, unidos para hacer una jaula aproximadamente esférica con tres pentágonos juntándose en cada vértice. Unamos ahora cada cara del dodecaedro con la cara opuesta... salvo que haya un giro. Literalmente. Cada cara tiene que rotarse un ángulo apropiado antes de unirla a la opuesta. El ángulo es el más pequeño que alinea las caras correspondientes, que es de 36 °. Podemos considerar esta regla como una versión elaborada de la regla de la banda de Möbius: girar un lado 180 ° y luego unirlo con el lado opuesto.

Ese es el espacio. Examinemos ahora el invariante. No me estoy yendo por las ramas: necesitamos todo esto para entender la conjetura de Poincaré.

Poincaré llamó a su nuevo invariante el grupo fundamental. Hoy seguimos utilizando ese nombre, pero también lo conocemos como el (primer) grupo de homotopía. La homotopía es una construcción geométrica que puede realizarse enteramente dentro del espacio, y proporciona información sobre el tipo topológico de dicho espacio. Lo hace utilizando una estructura algebraica abstracta conocida como un grupo. Un grupo es una colección de objetos matemáticos tal que dos cualesquiera de ellos pueden combinarse para dar otro objeto del grupo. Esta ley de composición —a menudo llamada multiplicación o suma, aunque no es la habitual operación aritmética con dicho nombre— debe satisfacer unas pocas condiciones simples y naturales. Si llamamos suma a esta operación, las condiciones principales son:

- El grupo contiene un elemento que se comporta como cero: si se suma a cualquier objeto del grupo, el resultado es el mismo objeto.
- Todo miembro del grupo tiene un negativo en el grupo: la suma de los dos da cero.
- Si se suman tres miembros del grupo, no importa qué dos se suman primero. Es decir, $(a + b) + c = a + (b + c)$. Esto se llama ley asociativa.

La ley algebraica que *no* se impone (aunque a veces también se cumple) es la ley conmutativa $a + b = b + a$[72].

[72] El convenio usual es hablar de suma y utilizar la notación $a + b$ cuando la ley de conmutación es válida, pero hablar de multiplicación y utilizar la notación ab cuando podría no

El grupo fundamental de Poincaré es una especie de esqueleto simplificado del espacio. Es un invariante topológico: espacios topológicamente equivalentes tienen el mismo grupo fundamental. Para hacernos una idea útil, y muy posiblemente reconstruir parte de la motivación de Poincaré, veamos cómo funciona en el caso de un círculo, tomando prestada una imagen que se remonta a Gauss. Imaginemos una hormiga cuyo universo entero es el círculo. ¿Cómo puede descubrir qué forma tiene su universo? ¿Puede distinguir el círculo de, pongamos por caso, una recta? Tengamos en cuenta que a la hormiga no se le permite salir fuera de su universo, mirarlo y ver que es circular. Todo lo que puede hacer es caminar dentro de su universo, cualquiera que pueda ser. En particular, la hormiga no se da cuenta de que su universo es curvo, porque su versión de un rayo de luz también está confinada en el círculo. Ignoremos cuestiones prácticas tales como que los objetos tengan que pasar unos a través de otros; esta va a ser una analogía muy vaga.

La hormiga tiene varias maneras de descubrir la forma de su universo. Me centraré en un método que se generaliza a cualquier espacio topológico. Para los fines de esta discusión, la hormiga es un punto. Vive en una parada de autobús que también es un punto. Cada día sale de su casa, toma el autobús (que, por supuesto, es un punto) y termina volviendo a casa. El trayecto más sencillo es el del autobús número 0, que simplemente se sitúa en la parada y no va a ninguna parte. Para una excursión más interesante, la hormiga toma el autobús número 1, que da la vuelta al universo

serlo. He ignorado aquí este convenio porque no es un libro de texto sobre teoría de grupos, y «suma» parece más natural.

exactamente una vez en sentido contrario a las agujas del reloj y se detiene cuando vuelve a casa. El autobús número 2 da dos vueltas, el número 3 da tres vueltas y así sucesivamente: un autobús en sentido contrario a las agujas del reloj por cada entero positivo. También hay autobuses negativos, que van en sentido contrario. El autobús número -1 da una vuelta en sentido de las agujas del reloj, el número -2 da dos vueltas en sentido de las agujas, y así sucesivamente.

La hormiga advierte pronto que dos viajes sucesivos en el autobús número 1 son, en esencia, lo mismo que un único viaje en el autobús número 2, y tres viajes en el autobús número 1 son esencialmente lo mismo que un único viaje en al autobús número 3. De modo similar, un viaje en el autobús número 5 seguido de un viaje en el autobús número 8 es esencialmente lo mismo que un viaje en el autobús número 13. De hecho, dados dos números positivos cualesquiera, un viaje en el autobús con el primer número seguido de un viaje en el autobús con el segundo número se reduce a un viaje en el autobús cuyo número es la suma de ambos.

El siguiente paso es más sutil. La misma relación es *casi* válida para autobuses cuyos números son negativos o cero. Un viaje en el autobús número 0, seguido de un viaje en el autobús número 1, es muy similar a un viaje en el autobús número 1. Sin embargo, hay una pequeña diferencia. En el viaje 0 + 1, el autobús 0 espera un tiempo en la salida, lo que no sucede en un único viaje en el autobús 1. Así que introducimos una noción con el imponente nombre de homotopía («el mismo lugar» en griego). Dos lazos son

homotópicos si uno puede ser deformado de forma continua hasta dar el otro. Si permitimos que los itinerarios de los autobuses se cambien por homotopías, podemos contraer poco a poco el tiempo que pasa la hormiga sentada en la parada del autobús número 0, hasta que el período estacionario desaparece. Ahora la diferencia entre el viaje 0 + 1 y el viaje 1 ha desaparecido, de modo que «salvo homotopía» el resultado es precisamente un viaje en el autobús número 1. Es decir, la ecuación 0 + 1 = 1 para números de autobús sigue siendo válida, no para viajes sino para clases de homotopía de viajes.

¿Qué pasa con un viaje en el autobús número 1 seguido de un viaje en el autobús número -1? Nos gustaría que esto fuera un viaje en el autobús número 0, pero no lo es. Recorre el trayecto en sentido contrario a las agujas del reloj, y luego vuelve a recorrerlo en sentido de las agujas. Esto es claramente diferente de pasar todo el viaje sentada en la parada del autobús número 0. De modo que 1 + (-1), es decir, 1 - 1 no es igual a 0. Pero una vez más la homotopía viene al rescate: la combinación de autobuses 1 y -1 es homotópica al mismo viaje total que el autobús 0. Para ver por qué, supongamos que la hormiga sigue en automóvil la ruta combinada de los autobuses 1 y -1, pero poco antes de completar todo el camino hasta la parada del autobús, invierte la dirección y vuelve a casa. Este viaje es muy parecido al doble viaje en autobús: solo difiere en una minúscula parte del recorrido. Así que el doble viaje en autobús original se ha «contraído» de forma continua a un viaje en automóvil ligeramente más corto. Ahora la hormiga puede contraer el viaje

otra vez, volviendo atrás un poco antes. Puede seguir contrayendo el recorrido, volviendo atrás un poco antes cada vez, hasta que finalmente todo lo que hace es sentarse en un automóvil aparcado en la parada de autobús, sin ir a ninguna parte. Este proceso de contracción es también una homotopía, y muestra que un viaje en el autobús número 1 seguido de un viaje en el autobús número -1 es homotópico a un viaje en el autobús número 0. Es decir, 1 + (-1) = 0 para clases de homotopía de viajes.

Es ahora sencillo, para un algebrista, demostrar que un viaje en cualquier autobús, seguido de un viaje en un segundo autobús, es homotópico a un viaje en el autobús que se obtiene sumando los dos números de autobús. Esto es cierto para autobuses positivos, autobuses negativos y el autobús 0. De modo que si sumamos viajes —bueno, clases se homotopía de viajes de autobús— obtenemos un grupo. De hecho, es un grupo muy familiar. Sus elementos son los enteros (números de autobús) y su operación es la suma. Su símbolo convencional es $\mathbb{Z}$, del alemán *Zahl* («entero»).

Un trabajo mucho más difícil demuestra que en un universo circular, *cualquier* viaje en automóvil —incluso si incluye muchos retrocesos o idas y venidas sobre el mismo tramo de carretera— es homotópico a uno de los viajes en autobús estándar. Además, los viajes en autobús con números diferentes no son homotópicos. La demostración requiere cierta técnica; la idea básica es el índice de Gauss. Este cuenta el número total de veces que el recorrido gira

alrededor del círculo en el sentido contrario a las agujas del reloj[73]. Dice a qué ruta de autobús es homotópico el viaje.

Una vez que se rellenan los detalles, esta descripción demuestra que el grupo fundamental de un círculo es el mismo que el grupo $\mathbb{Z}$ de enteros bajo suma. Para sumar viajes, simplemente se suman sus índices. La hormiga podría utilizar este invariante topológico para distinguir un universo circular de, pongamos por caso, una recta infinita. En una recta, cualquier viaje, por muchos vaivenes que tenga, debe en alguna etapa alcanzar una distancia máxima a casa. Ahora podemos contraer el viaje de forma continua contrayendo todas las distancias a casa en la misma cantidad, primero al 99 por 100, luego al 98 por 100, y así sucesivamente. De modo que sobre una recta, *cualquier* viaje es homotópico a cero: quedarse en casa. El grupo fundamental de la recta tiene un solo elemento: 0. Sus propiedades algebraicas son triviales: 0 + 0 = 0. Por eso se le llama el grupo trivial, y puesto que difiere del grupo de todos los enteros, la hormiga puede distinguir la diferencia entre vivir en una recta y vivir en un círculo.

Como dije, existen otros métodos, pero así es como la hormiga puede hacerlo utilizando el grupo fundamental de Poincaré.

Ahora subimos la apuesta. Supongamos que la hormiga vive en una superficie. De nuevo, ese es su universo entero; no puede salir de él y echar una mirada para ver en qué tipo de superficie habita.

[73] Empecemos la cuenta en cero. Cada vez que pasamos por la parada del autobús en sentido contrario a las agujas del reloj, incrementamos la cuenta en 1; cada vez que pasamos en el sentido de las agujas, disminuimos la cuenta en 1. Al final del viaje sumamos 1 si llegamos en sentido contrario a las agujas, y restamos 1 si llegamos en el sentido de las agujas. La cuenta final es el número total de veces que dimos vueltas al círculo, medidas en el sentido contrario a las agujas del reloj.

¿Puede calcular la topología de su universo? En particular, ¿puede distinguir la diferencia entre una esfera y un toro? De nuevo la respuesta es «sí», y el método es el mismo que para un universo circular: subir a un autobús y hacer viajes que empiezan y acaban en casa. Para sumar viajes, se realizan uno detrás de otro. El viaje cero es «quedarse en casa», el inverso de un viaje es el mismo viaje en sentido contrario, y obtenemos un grupo con tal de que trabajemos con clases de homotopía de viajes. Este es el grupo fundamental de la superficie. Comparado con un universo circular hay más libertad para crear viajes y para deformarlos de forma continua en otros viajes, pero funciona la misma idea básica.

El grupo fundamental es de nuevo un invariante topológico, y la hormiga puede utilizarlo para descubrir si vive en una esfera o en un toro. Si su universo es una esfera, entonces, sea cual sea el viaje que haga la hormiga, puede deformarse poco a poco hasta el viaje cero: quedarse en casa. Esto no sucede si el universo es un toro. Algunos viajes pueden deformarse hasta cero, pero un viaje que da una vuelta a través del agujero central, como en la Figura 39 (*izquierda*), no puede deformarse así. Esta afirmación necesita demostración, pero puede darse. Hay viajes en autobús estándar en el toro, pero ahora los números de autobús son pares de enteros (*m*, *n*). El primer número *m* especifica cuántas vueltas da el viaje alrededor del agujero central; el segundo número *n* especifica cuántas vueltas da el viaje alrededor del toro. La Figura 39 (*derecha*) muestra el viaje (5, 2), que da cinco vueltas a través del agujero y dos vueltas alrededor del toro. Para sumar viajes, se suman los

números correspondientes; por ejemplo, (3, 6) + (2, 4) = (5, 10). El grupo fundamental del toro es el grupo $\mathbb{Z}^2$ de *pares* de enteros.

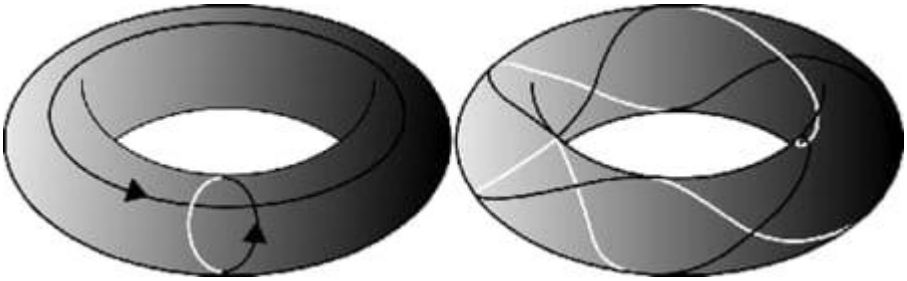


Figura 39. Izquierda: Viajes en autobús (1, 0) y (0, 1) en el toro. Derecha: Viaje en autobús (5, 2). Las líneas grises están en la parte trasera.

Cualquier espacio topológico tiene un grupo fundamental, definido exactamente de la misma forma utilizando viajes —conocidos con más propiedad como lazos— que empiezan y terminan en el mismo punto. Poincaré ideó el grupo fundamental para demostrar que su espacio dodecaédrico no es una 3-esfera, pese a tener exactamente los mismos invariantes topológicos. Su fórmula original se adapta muy bien al cálculo de su grupo fundamental. La receta más moderna «girar y pegar» se adapta todavía mejor. La respuesta resulta ser un grupo con 120 elementos relacionados con el dodecaedro. En contraste, el grupo fundamental de una 3-esfera tiene un solo elemento: el lazo cero. De modo que el espacio dodecaédrico no es topológicamente equivalente a una esfera, pese a tener la misma homología, y Poincaré había demostrado que su afirmación de 1900 era errónea.

Él pasó a especular sobre su nuevo invariante: ¿era el ingrediente que faltaba en una caracterización topológica de la 3-esfera? ¿Quizá cualquier espacio 3-dimensional con el mismo grupo fundamental que una 3-esfera —es decir, el grupo trivial— debe *ser* realmente una 3-esfera? Parafraseó esta sugerencia en forma negativa como una pregunta: «Consideremos una variedad [espacio topológico] 3-dimensional compacta *V* sin frontera. ¿Es posible que el grupo fundamental de *V* sea trivial, incluso si *V* no es [topológicamente] equivalente a la esfera 3-dimensional?». Él dejó la pregunta abierta, pero la creencia muy plausible de que la respuesta es la obvia —«no», cuando la pregunta se plantea de esta manera— llegó rápidamente a conocerse como la conjetura de Poincaré. Y con la misma rapidez llegó a ser una de las más enojosas preguntas abierta en topología.

«Grupo fundamental trivial» es otra manera de decir «todo lazo puede ser deformado de forma continua hasta un punto». No solo una 3-esfera posee esta propiedad; también posee una *n*-esfera análoga para cualquier dimensión *n*. Por ello podemos hacer la misma conjetura para una esfera de cualquier dimensión. Esta afirmación es la conjetura de Poincaré *n*-dimensional. Es verdadera cuando $n = 2$, por el teorema de clasificación para superficies. Y durante más de cincuenta años, eso fue lo más lejos que se pudo llegar.

En 1961 Stephen Smale tomó prestado un truco de la clasificación de superficies y lo aplicó en dimensiones más altas. Una manera de pensar en un toro *g*-agujereado es empezar con una esfera y añadir

g asas —igual que el asa de una taza o un jarra—. Smale generalizó esta construcción a cualquier número de dimensiones, y llamó al proceso descomposición en asas. Analizó cuántas asas podían modificarse sin cambiar la topología del espacio, y dedujo la conjetura de Poincaré en todas las dimensiones mayores o iguales que 7. Su demostración no era válida para dimensiones menores, pero otros matemáticos encontraron una manera de repararlo: John Stallings para dimensión 6 y Christopher Zeeman para dimensión 5. Sin embargo, un paso vital, conocido como truco de Whitney, fallaba para dimensiones 3 y 4 porque en estos espacios no hay lugar suficiente para realizar las maniobras requeridas, y nadie pudo encontrar sustitutos efectivos. Cundió una sensación general de que la topología para espacios de estas dos dimensiones podría ser inusual.

Esta opinión convencional sufrió una sacudida en 1982 cuando Michael Freedman descubrió una demostración para la conjetura de Poincaré 4-dimensional que no requería el truco de Whitney. Era extremadamente complicada, pero funcionaba. Así, después de cincuenta años de pequeños progresos y veinte años de actividad frenética, los topólogos habían despachado la conjetura de Poincaré en cualquier dimensión salvo en aquella por la que Poincaré había preguntado originalmente. Los éxitos eran impresionantes, pero los métodos utilizados para obtenerlos proporcionaban muy pocas intuiciones sobre el caso 3-dimensional. Se necesitaba una manera de pensar diferente.

Lo que finalmente rompió el punto muerto fue algo parecido a la lista tradicional de regalos de boda: algo viejo, algo nuevo, algo prestado... y, alterando un detalle, algo azul. La idea vieja era revisitar un área de la topología de la que se pensaba que, tras haber sido campo de una frenética actividad en espacios de dimensiones más altas, estaba agotada: la topología de superficies. La idea nueva era replantear la clasificación de superficies desde un punto de vista que al principio parecía completamente ajeno: la geometría clásica. La idea prestada era el flujo de Ricci, que tomó su motivación del formalismo matemático de la teoría de la relatividad general de Einstein. Y la idea azul era especulación «caída del cielo»: algunas sugerencias de gran alcance basadas en un toque de intuición y mucho de esperanza.

Recordemos que puede hacerse una lista de las superficies orientables sin frontera: cada una de ellas es topológicamente equivalente a un toro con cierto número de agujeros. Este número es el género de la superficie, y cuando es cero, la superficie es una esfera sin asas, es decir, una esfera. La propia palabra nos recuerda que entre todas las esferas topológicas hay una superficie que representa a todas las demás como arquetipo; a saber, la esfera unidad en el espacio euclídeo. Olvidemos, por un segundo, todo eso de la lámina elástica. Volveremos a ponerlo en un momento. Concentrémonos en el viejo espacio euclídeo. Tiene todo tipo de propiedades matemáticas extra, derivadas de la rigidez de la geometría euclídea. Primordial entre estas propiedades es la curvatura. La curvatura puede cuantificarse; en cada punto en una

geometría euclídea hay un número que mide cuán curvada está la superficie cerca de dicho punto. La esfera es la única superficie cerrada en el espacio euclídeo cuya curvatura es la misma en todo punto, y es positiva.

Esto es extraño, porque curvatura constante no es una propiedad topológica. Más extraño todavía: la esfera no está sola. También hay una superficie geométrica estándar que destaca como el toro arquetípico: se empieza con un cuadrado en el plano y se identifican lados opuestos (véase la Figura 12 del capítulo 4). Cuando dibujamos el resultado en el espacio 3-dimensional, enrollando el cuadrado para hacer que sus lados de encuentren, el resultado parece curvado. Pero desde un punto de vista intrínseco, podemos trabajar con el cuadrado más las reglas de unión. Un cuadrado tiene una estructura geométrica natural: es una región en el plano euclídeo. El plano tiene también curvatura constante, pero ahora la constante es *cero*. Un toro con esta geometría particular tiene también curvatura cero, por lo que es llamado el *toro plano*. El nombre puede parecer una contradicción en los términos, pero para una hormiga que viva en un toro plano, que lleva una regla y un transportador para medir longitudes y ángulos, la geometría local sería idéntica a la del plano.

Los geómetras del siglo XVIII, que trataban de entender el axioma de Euclides sobre la existencia de paralelas, se propusieron deducir dicho axioma a partir de las demás hipótesis básicas de Euclides. No lo consiguieron por más que lo intentaran y terminaron dándose cuenta de que tal deducción no es posible. Hay tres tipos de

geometría diferentes, cada uno de los cuales satisface todas las condiciones que requiere Euclides salvo el axioma de las paralelas. Dichas geometrías se llaman euclídea (la plana, donde el axioma de las paralelas es válido), elíptica (geometría en la superficie de una esfera, con unos pocos elementos accesorios, donde dos rectas cualesquiera se encuentran y las paralelas no existen), e hiperbólica (donde algunas rectas no se encuentran, y las paralelas no son únicas). Además, los matemáticos clásicos interpretaron estas geometrías como la geometría de espacios curvos. La geometría euclídea corresponde a curvatura cero, la geometría elíptica/esférica corresponde a curvatura positiva constante, y la geometría hiperbólica corresponde a curvatura negativa constante.

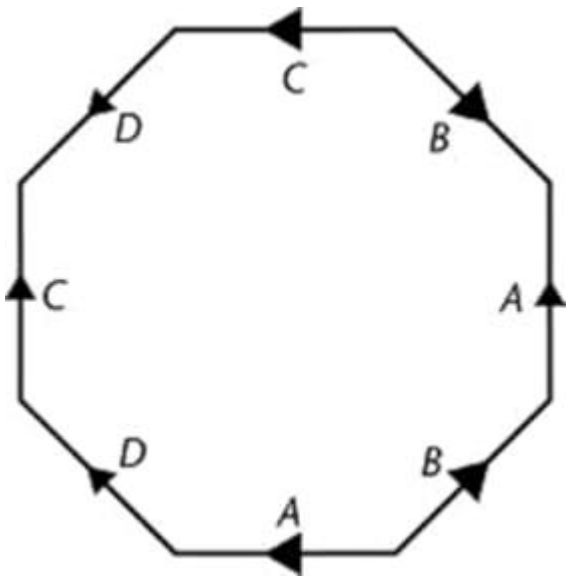


Figura 40. Construcción de un toro 2-agujereado a partir de un octógono identificando lados en pares (AA, BB, CC, DD).

Acabamos de ver cómo se obtienen las dos primeras de estas geometrías: se dan en la esfera y en el toro plano. En términos del teorema de clasificación, son toros g-agujereados para $g = 0$ y 1. La única que falta es la geometría hiperbólica. ¿Tiene todo toro g-agujereado una estructura geométrica natural, basada en tomar un polígono en el espacio hiperbólico e identificar algunos de sus lados? La respuesta es sorprendente: es «sí» para *cualquier* valor de g mayor o igual que 2. La Figura 40 muestra un ejemplo para $g = 2$ basado en un octógono. Saltaré la geometría hiperbólica y la identificación de esta superficie como un 2-toro, pero pueden explicarse. Aparecen diferentes valores de g si tomamos polígonos diferentes, pero se dan todos los g. En la jerga, un toro con dos o más agujeros tiene una estructura hiperbólica natural. De modo que ahora podemos reinterpretar la lista de superficies estándar:

- Esfera, $g = 0$: geometría elíptica.
- Toro, $g = 1$: geometría euclídea.
- Toro g-agujereado, $g = 2, 3, 4,\ldots$: geometría hiperbólica.

Puede parecer que hemos tirado al niño con el agua de bañarlo, porque se supone que la topología trata de la geometría de la lámina elástica, no la geometría rígida. La geometría rígida se utiliza aquí solamente para *definir* las superficies estándar. Proporciona descripciones simples, que resultan tener estructura rígida extra. Relajemos ahora la rigidez y permitamos que el espacio se haga algo parecido a goma. Deformémoslo de maneras que prohíbe la rigidez. Ahora obtenemos superficies que son topológicamente equivalentes

a las estándar, pero no son equivalentes por movimientos rígidos. El teorema de clasificación nos dice que toda superficie topológica puede obtenerse de esta manera.

Los topólogos eran conscientes de este vínculo entre geometría y el teorema de clasificación de superficies, pero parecía una divertida coincidencia, sin duda una consecuencia de las posibilidades bastante limitadas en dos dimensiones. Todos sabían que el caso 3-dimensional era mucho más rico, y en particular que los espacios de curvatura constante no agotaban las posibilidades. Fue necesario que uno de los mejores geómetras del mundo, William Thurston, se diera cuenta de que la geometría rígida también podría ser relevante para la topología 3-dimensional. Había ya algunos indicios: la 3-esfera de Poincaré tiene una geometría elíptica/esférica natural, que procede de su definición. Aunque un dodecaedro estándar vive en el espacio euclídeo, el ángulo entre caras adyacentes es menor de 120 °, de modo que tres de estos ángulos no cubren un círculo completo. Para remediarlo, tenemos que inflar el dodecaedro de modo que sus caras se abomben ligeramente: esto hace natural la geometría esférica, no la euclídea. De modo análogo, los triángulos en una esfera también se abomban. El 3-toro, obtenido identificando caras opuestas de un cubo, tiene una geometría plana —es decir, euclídea— igual que su análogo 2-dimensional. Max Dehn y otros habían descubierto algunos espacios topológicos 3-dimensionales con geometrías hiperbólicas naturales.

Thurston empezó a ver indicios de una teoría general, pero se necesitaban dos innovaciones para hacerla siquiera remotamente

plausible. Primera: había que ampliar el rango de geometrías 3-dimensionales. Thurston estableció condiciones razonables y demostró que las satisfacían exactamente ocho geometrías. Tres de ellas son las clásicas: geometría esférica, euclídea e hiperbólica. Dos más son como cilindros: plana en una dirección, curva en las otras dos direcciones. La parte curvada está o bien curvada positivamente, la 2-esfera, o bien curvada negativamente, el plano hiperbólico. Finalmente, hay otras tres geometrías bastante técnicas.

Segunda: algunos espacios 3-dimensionales no soportaban ninguna de las ocho geometrías. La respuesta estaba en cortar el espacio en piezas. Una pieza podría tener una estructura geométrica esférica, otra podría tener una estructura hiperbólica, y así de modo sucesivo. Para ser útil, el corte tenía que hacerse de una manera muy fuertemente restringida, de modo que el reensamblaje de las piezas llevara información útil. Las buenas noticias eran que en muchos ejemplos esto resultó ser posible. En 1982, en un gran salto de imaginación, Thurston enunció su conjetura de geometrización: *todo* espacio 3-dimensional puede cortarse en piezas, de manera esencialmente única, cada una de las cuales tiene una estructura geométrica natural correspondiente a una de las ocho geometrías posibles. También demostró que si su conjetura de geometrización fuera cierta, entonces la conjetura de Poincaré sería una simple consecuencia.

Mientras tanto estaba emergiendo una segunda línea de ataque, también geométrica, asimismo basada en la curvatura, pero

procedente de un área muy diferente: la física matemática. Gauss, Riemann y una escuela de geómetras italianos habían desarrollado una teoría general de espacios curvos, llamados variedades, con un concepto de distancia que ampliaba enormemente la geometría euclídea y la geometría no euclídea clásica. La curvatura ya no tenía que ser constante: podía variar suavemente de un punto a otro. Una forma como la de un hueso de perro, por ejemplo, está curvada en sentido positivo en cada extremo pero curvada en sentido negativo en medio, y la cantidad de curvatura varía suavemente de una región a la próxima. La curvatura se cuantifica utilizando artefactos matemáticos conocidos como tensores. Alrededor de 1915 Albert Einstein se dio cuenta de que los tensores de curvatura eran exactamente lo que necesitaba para ampliar su teoría de la relatividad especial, que trataba del espacio y el tiempo, hasta la relatividad general, que también incluía la gravedad. En esta teoría el campo gravitatorio es representado como la curvatura del espacio, y las ecuaciones de campo de Einstein describen cómo cambia la medida de curvatura asociada, el tensor de curvatura, en respuesta a la distribución de materia. En efecto, la curvatura del espacio *fluye* con el paso del tiempo; el universo o alguna parte del mismo cambian su forma espontáneamente.

Richard Hamilton, un especialista en geometría riemanniana, comprendió que el mismo truco podría aplicarse de manera más general, y eso podría llevar a una demostración de la conjetura de Poincaré. La idea era trabajar con una de las medidas de curvatura más simples, llamada curvatura de Ricci en referencia al geómetra

italiano Gregorio Ricci-Curbastro. Hamilton escribió una ecuación que especificaba cómo debería cambiar la curvatura de Ricci con el tiempo: el flujo de Ricci. La ecuación estaba establecida de modo que la curvatura debería redistribuirse gradualmente de una manera lo más uniforme posible. Esto es un poco como el gato bajo una alfombra en el capítulo 4, pero ahora, incluso si el gato no puede escapar, puede extenderse en una capa uniforme. (Aquí hay, en esencia, un gato topológico).

Por ejemplo, en el caso 2-dimensional partimos de una superficie en forma de pera (véase Figura 41). Esta tiene una región en un extremo que está fuertemente y en sentido positivo curvada; una región en el otro extremo, más grueso, que está también curvada en sentido positivo, aunque no tan fuertemente; y una zona en medio donde la curvatura es negativa. El flujo de Ricci transporta en efecto curvatura desde el extremo fuertemente curvado (y en menor medida desde el otro extremo) a la zona curvada en sentido negativo, hasta que toda la curvatura negativa ha sido absorbida. En esa etapa el resultado es una superficie abombada con curvatura positiva en todo lugar. El flujo de Ricci continúa redistribuyendo curvatura, llevándosela de las regiones altamente curvadas y pasándola a las regiones menos curvadas. Conforme crece el tiempo, la superficie se acerca cada vez más a una que tiene curvatura positiva constante, es decir, una esfera euclídea. La topología sigue siendo la misma, incluso si cambia la forma detallada, de modo que podemos demostrar que la superficie

original con forma de pera es topológicamente equivalente a una esfera siguiendo el flujo de Ricci.

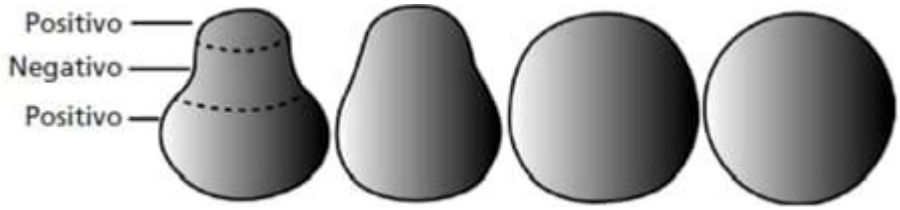


Figura 41. Cómo el flujo de Ricci convierte una pera en una esfera.

En este ejemplo el tipo topológico de la superficie era obvio de entrada, pero la misma estrategia general funciona para cualquier variedad. Partamos de una forma complicada y sigamos el flujo de Ricci. Conforme pasa el tiempo la curvatura se redistribuye de manera más uniforme y la forma se hace más simple. Finalmente, uno debería terminar con la forma más simple que tiene la misma topología que la variedad original, cualquiera que pueda ser. En 1981 Hamilton demostró que esta estrategia funciona en dos dimensiones, lo que proporcionaba una nueva demostración del teorema de clasificación para superficies.

Él hizo también progresos importantes sobre la estrategia análoga para variedades 3-dimensionales, pero ahora había un serio obstáculo. En dos dimensiones toda superficie se simplifica de forma automática siguiendo el flujo de Ricci. Lo mismo es cierto en tres dimensiones si la variedad inicial tiene curvatura estrictamente positiva en todo punto: nunca curvatura cero o negativa. Por desgracia, si hay puntos en los que la curvatura es cero, y a menudo los hay, el espacio puede entrelazarse consigo mismo a

medida que fluye. Esto crea singularidades: lugares donde la variedad deja de ser suave. En dichos puntos la ecuación para el flujo de Ricci deja de ser válida y la redistribución de curvatura tiene que detenerse. La forma natural de evitar este obstáculo es entender a qué se parecen las singularidades y rediseñar la variedad —quizá cortándola en piezas— de modo que se pueda dar una nueva salida al flujo de Ricci. Siempre que se tenga suficiente control sobre cómo se relaciona la topología de la variedad remodelada con la de la original, esta estrategia modificada puede ser acertada. Por desgracia, Hamilton advirtió también que en el caso de espacios 3-dimensionales las singularidades en el flujo de Ricci pueden ser muy complicadas, demasiado complicadas, al parecer, para utilizar un truco de ese tipo. El flujo de Ricci se convirtió rápidamente en una técnica estándar en geometría, pero estaba lejos de demostrar la conjetura de Poincaré.

En 2000 los matemáticos todavía no habían resuelto la conjetura, y su importancia tuvo un reconocimiento mucho mayor cuando se hizo de ella uno de los siete problemas del milenio. Para entonces también había quedado claro que si de algún modo pudiera hacerse que la idea de Hamilton funcionase con suficiente generalidad, no solo implicaría la conjetura de Poincaré sino que también demostraría la conjetura de geometrización de Thurston. El premio era rutilante, pero permanecía tentadoramente fuera del alcance.

Las matemáticas son como las demás ramas de la ciencia: para que la investigación sea aceptada como correcta tiene que estar publicada, y para que eso suceda tiene que superar la evaluación

por los pares. Expertos en el campo tienen que leer el artículo cuidadosamente, comprobar que los razonamientos son correctos y asegurarse de que no hay errores en los cálculos. Este proceso puede llevar mucho tiempo en el caso de un trabajo matemático importante y complicado. Como se mencionó en el capítulo 1, el remedio solía ser una prepublicación, pero en nuestros días hay una página web estándar, los arXiv («archivos»), donde pueden colocarse prepublicaciones en formato electrónico, sometidos a un proceso de revisión parcial y un procedimiento de aceptación para evitar disparates. Hoy, la mayoría de los investigadores encuentran por primera vez nuevos resultados en los arXiv o en la propia página web del autor.

En 2002 Grigori Perelman colocó en los arXiv una prepublicación sobre el flujo de Ricci. Hacía una afirmación extraordinaria: el flujo es de tipo gradiente. Es decir, hay una dirección «descendente» bien definida, una simple cantidad numérica asociada con la forma de la variedad, y la variedad fluye pendiente abajo en el sentido de que esta cantidad siempre decrece conforme pasa el tiempo. Es análoga a la altura en un paisaje, y proporciona una medida cuantitativa de lo que significa «simplificar» una variedad. Los flujos de tipo gradiente son bastante restringidos: pueden dar vueltas y vueltas en círculo o comportarse de forma caótica. Nadie parece haber sospechado que el flujo de Ricci fuera tan dócil. Pero Perelman no solo hacía la afirmación: él la demostraba. Terminaba esbozando un argumento que corroboraría la conjetura de geometrización de Thurston —que implica la conjetura de Poincaré pero va mucho más

lejos— y prometía más detalles en una posterior aportación en los arXiv. Durante los ocho meses siguientes colocó otros dos artículos que continuaban el trabajo y contenían muchos de los detalles prometidos.

El primero causó un revuelo. Perelman estaba afirmando que había llevado a cabo el programa de Hamilton completo, utilizando el flujo de Ricci para simplificar una variedad 3-dimensional y demostrar que el resultado era exactamente el que Thurston había predicho. Los otros dos añadían más peso a la sensación de que Perelman sabía de lo que estaba hablando, y que sus ideas iban mucho más allá de esbozar una estrategia plausible con la consabida laguna lógica o una hipótesis no demostrada. El escepticismo habitual de la comunidad matemática respecto a las pretensiones de haber resuelto un gran problema cambió; la impresión general era que muy bien él podría haberlo conseguido.

Sin embargo, el diablo está en los detalles, y en matemáticas los detalles pueden ser realmente infernales. El trabajo tenía que ser comprobado, en extensión y en profundidad, por personas entendidas en las áreas implicadas y que fueran conscientes de las trampas potenciales. Y eso no era sencillo, porque Perelman había combinado al menos cuatro áreas muy diferentes de las matemáticas y la física matemática, y pocas personas entendían más de una o dos de ellas. Decidir si su demostración era correcta requeriría mucho trabajo en equipo y mucho esfuerzo. Además, las prepublicaciones en los arXiv no incluían todos los detalles con el nivel que es habitual en un artículo impreso. Estaban escritos de

forma muy clara para lo que son las prepublicaciones, pero no siempre estaban los puntos sobre las íes. Por ello los expertos tenían que reconstruir parte del pensamiento de Perelman —y él había estado profundamente inmerso en el trabajo durante años.

Todo ello llevó tiempo. Perelman dio conferencias sobre su demostración y respondió a correos electrónicos que cuestionaban varios pasos. Cada vez que alguien encontraba lo que parecía una laguna, él respondía al momento con más explicaciones y la llenaba. Los signos eran alentadores. Pero nadie iba a arriesgar su reputación afirmando en público que Perelman había demostrado la conjetura de Poincaré, y mucho menos la difícil conjetura de geometrización, hasta que estuvieran seguros de que no había errores en la demostración. De modo que pese a la opinión general favorable sobre el trabajo de Perelman, la aceptación pública fue inicialmente aplazada. Esto era inevitable, pero también desafortunado, porque conforme aumentaba la espera, la irritación de Perelman crecía frente a lo que le parecía pasividad. Él sabía que su demostración era correcta. La entendía tan bien que no podía ver por qué otros estaban teniendo dificultades. Declinó exponer el trabajo con más detalle o enviarlo a una revista. Por lo que a él concernía, era cosa hecha, y las prepublicaciones en los arXiv contenían todo lo que se requería. Dejó de responder a preguntas sobre los detalles que presuntamente faltaban. Para él, no faltaban. Vamos muchachos, podéis imaginarlo sin más ayuda por mi parte. No es tan difícil.

Algunos informes han sugerido que a este respecto la comunidad matemática fue injusta con Perelman. Pero esto malinterpreta cómo funciona la comunidad matemática cuando supuestamente se ha resuelto un gran problema. Hubiera sido irresponsable darle palmadas en la espalda, decir «¡bien hecho!» e ignorar los pasos que faltaban en sus prepublicaciones. Era perfectamente lógico, de hecho inevitable, pedirle que preparara tratamientos más extensos, adecuados para publicación. En un problema de esta importancia, un trabajo apresurado es peligroso e inaceptable. Los expertos dejaban su trabajo para dedicar mucho tiempo a la demostración de Perelman, y mantenían a raya su escepticismo natural en una medida inusual. Su tratamiento fue en cualquier caso *más* favorable que el habitual. Y finalmente, cuando se completó este proceso, su trabajo fue aceptado como correcto.

Para entonces, no obstante, Perelman había perdido la paciencia. Tal vez no ayudó el hecho de que él hubiera resuelto un problema tan importante que no había otro que se le pareciese. Era como un montañero que había escalado el Everest solo sin oxígeno. No quedaban desafíos comparables. La publicidad en los medios de comunicación le repelía: él buscaba la aceptación de sus pares, no de presentadores de televisión. Por ello no puede sorprender mucho que cuando sus pares acordaron finalmente que él estaba en lo cierto, y le ofrecieron una medalla Fields y el premio Clay, él no quiso saber nada.

La demostración de Perelman es profunda y elegante, y abre un nuevo mundo a la topología. Implementa el programa de flujo de

Ricci de Hamilton al encontrar maneras más inteligentes para evitar la aparición de singularidades. Una es cambiar las escalas de espacio y tiempo para deshacerse de las singularidades. Cuando esta aproximación falla, se dice que la singularidad colapsa. En tales casos, él analiza en detalle la geometría del flujo de Ricci, clasificando cómo puede ocurrir un colapso. En efecto, el espacio saca tentáculos cada vez más finos, quizá en profusión, como las ramas de un árbol. Cada vez que un tentáculo está próximo a colapsar, puede ser cortado, y el extremo puntiagudo fuertemente curvado puede ser seccionado y reemplazado por un capuchón suave. Para algunos de estos tentáculos el flujo de Ricci llega a detenerse: si lo hace, se le deja solo. Si no, el flujo de Ricci puede ser reiniciado. De modo que algunos tentáculos terminan en capuchones suaves, y otros son temporalmente interrumpidos, pero siguen fluyendo.

Este procedimiento corta-y-pega de capuchones divide el espacio de forma muy parecida a la disección de Thurston en piezas, cada una de ellas con una de sus ocho geometrías, y los dos procedimientos dan resultados más o menos idénticos. Un punto técnico es vital: las operaciones de colocar capuchones no se acumulan a un ritmo cada vez mayor que haga que infinitas de ellas ocurran en un tiempo finito. Esta es una de las partes más complicadas de la demostración.

Algunos comentaristas han criticado a la comunidad matemática por tratar injustamente a Perelman. Nadie debería ser inmune a las críticas, y hubo algunos incidentes que muy bien podrían ser

clasificados como injustos o poco considerados en otros aspectos, pero la comunidad matemática reaccionó rápida y positivamente al trabajo de Perelman. También reaccionó con cautela, lo que es absolutamente estándar en matemáticas y ciencia, por excelentes razones. El inevitable foco de la publicidad, aumentado por el premio del millón de dólares, tuvo un impacto sobre todo el mundo, incluido Perelman.

Desde el primer envío de Perelman a los arXiv en noviembre de 2002 hasta el anuncio en marzo de 2010 de que se le había concedido el premio Clay pasaron ocho años. Eso suena como un largo retraso, un retraso quizá irrazonable. Sin embargo, el primer envío trataba solo parte del problema. Casi todo lo demás fue enviado en marzo de 2003. Para septiembre de 2004, dieciocho meses después de su segundo envío, las comunidades del flujo de Ricci y de la topología ya habían revisado la demostración —un proceso que había empezado pocos días después del primer envío— y los principales expertos anunciaron que «entendían la demostración». Habían encontrado errores, habían encontrado lagunas, pero estaban convencidos de que todas podían corregirse. Dieciocho meses es realmente muy rápido cuando está en juego algo tan importante.

Luego, en 2005, la Unión Matemática Internacional se dirigió a Perelman y le ofreció una medalla Fields, el máximo honor de la disciplina, que sería concedida en el Congreso Internacional de Matemáticos de 2006[ii]. El ICM se celebra cada cuatro años, así que esta era la primera oportunidad de reconocer su trabajo de esta manera. Puesto que quedaban algunas dudas sobre la demostración

completa de la conjetura de Poincaré —seguían apareciendo errores—, la medalla fue oficialmente concedida por avances en la comprensión del flujo de Ricci, la parte de las prepublicaciones de Perelman que entonces se consideraba libre de errores.

Las condiciones para la concesión del premio están establecidas en la página web del Instituto Clay. En particular, una solución propuesta tiene que estar publicada en una revista con revisión por pares y además ser aceptada por la comunidad matemática dos años después. Después de eso, un comité asesor especial examina la materia y recomienda si conceder o no el premio. Perelman no había satisfecho la primera condición, y no parece probable que lo hiciera alguna vez. En su opinión, las prepublicaciones en arXiv bastaban. Sin embargo, el Instituto Clay soslayó ese requisito e inició la espera estatutaria de dos años para ver si aparecían otros errores u otras cuestiones. Ese plazo terminó en 2008, después de lo cual tenían que seguirse los protocolos del Instituto, cuidadosamente estructurados para evitar la concesión prematura del premio.

Es cierto que algunos expertos fueron lentos en expresar su creencia de que la demostración era correcta. La razón es sencilla: de hecho, estaban inseguros. No es una gran exageración decir que la única persona capaz de captar rápidamente la demostración de Perelman era otro Perelman. No se puede leer una demostración matemática como un músico lee a primera vista una partitura. Uno tiene que convencerse de que todo tiene sentido. Cada vez que el argumento se hace muy complicado, uno sabe que hay una seria

posibilidad de error. Lo mismo sucede cuando las ideas son demasiado simples; más de una presunta demostración ha sido víctima de una afirmación tan evidente que no parecía necesaria ninguna demostración. Hasta que los expertos no estuvieran por completo seguros de que la demostración era básicamente correcta —momento en el cual dieron a Perelman todo el crédito *pese a* las lagunas y errores que quedaban—, era razonable suspender el juicio. Pensemos en todo el revuelo que causó el finalmente desacreditado trabajo sobre la fusión fría. La cautela es la respuesta profesional correcta, y aquí se aplica el cliché: afirmaciones extraordinarias requieren pruebas extraordinarias.

¿Por qué rechazó Perelman la medalla Fields y declinó el premio Clay? Solo él lo sabe, pero no estaba interesado en ese tipo de reconocimiento y así lo dijo de modo reiterado. Ya había rechazado premios menores. Dejó claro de entrada que no quería publicidad prematura; irónicamente, esta es la misma razón por la que los expertos eran comprensiblemente reacios a pronunciarse demasiado pronto. Para ser realistas, no había la más mínima posibilidad de que los medios de comunicación *no* advirtieran este trabajo. Durante años la comunidad matemática ha estado haciendo un gran esfuerzo para que los periódicos, la radio y la televisión se interesen en la materia. No tiene mucho sentido quejarse cuando este esfuerzo tiene éxito, o esperar que los medios ignoren la historia matemática más candente desde el último teorema de Fermat. Pero Perelman no lo veía así y se encerró en su concha. Hay una oferta sobre la mesa para que el dinero del premio se dedique a

fines educativos u otros si él está de acuerdo. Hasta ahora, no ha habido ninguna respuesta por su parte.

Capítulo 11
No todos pueden ser fáciles
El problema P/NP

Hoy día los matemáticos utilizan los ordenadores de manera rutinaria para resolver problemas, incluso grandes problemas. Los ordenadores son buenos para hacer aritmética, pero las matemáticas van mucho más allá de las meras «cuentas», de modo que poner un problema en un ordenador no es nada sencillo. A menudo la parte más difícil del trabajo es convertir el problema en otro problema que un cálculo por ordenador pueda resolver, e incluso entonces el ordenador puede tener dificultades. Muchos de los grandes problemas que han sido resueltos recientemente incluyen poco o ningún trabajo con un ordenador. El último teorema de Fermat y la conjetura de Poincaré son ejemplos.

Cuando los ordenadores han sido utilizados para resolver grandes problemas, como el teorema de los cuatro colores o la conjetura de Kepler, el ordenador desempeña efectivamente el papel de sirviente. Pero a veces los papeles se invierten y son las matemáticas las que actúan como sirvientes de la ciencia de la computación. La mayor parte de los trabajos iniciales sobre el diseño de ordenadores hacían buen uso de intuiciones matemáticas, por ejemplo la conexión entre el álgebra booleana —una formulación algebraica de la lógica— y los circuitos con interruptores, desarrollados en particular por el ingeniero Claude Shannon, el inventor de la teoría de la información. Hoy, aspectos tanto prácticos como teóricos de los

ordenadores se basan en el uso extensivo de matemáticas de áreas muy diferentes.

Uno de los problemas Clay del milenio se sitúa en la frontera entre las matemáticas y la ciencia de la computación. Puede verse de ambas maneras: la ciencia de la computación como sirviente de las matemáticas y las matemáticas como sirvientes de la ciencia de la computación. Lo que ello requiere, y está ayudando a conseguir, es más equilibrado: una asociación. El problema trata de los algoritmos de computación, los esqueletos matemáticos de los que están hechos los programas para computación. El concepto crucial aquí es cuán eficiente es el algoritmo: cuántos pasos computacionales necesita para obtener una respuesta para una cantidad dada de datos de entrada. En términos prácticos, esto nos dice cuánto tardará el ordenador en resolver un problema de un tamaño dado.

La palabra algoritmo se remonta a la Edad Media, cuando Muhammad ibn Mūsā al-Kwārizmī escribió uno de los primeros libros de álgebra. Previamente, Diofanto había introducido un elemento que asociamos con el álgebra: los símbolos. Sin embargo, él utilizaba los símbolos como abreviaturas, y sus métodos para resolver ecuaciones se presentaban mediante ejemplos concretos, aunque típicos. Donde ahora escribiríamos algo como «$x + a = y$, luego $x = y - a$», Diofanto escribiría «supongamos $x + 3 = 10$, entonces $x = 10 - 3 = 7$» y esperaría que sus lectores entendieran que la misma idea funcionaría si 3 y 10 fueran reemplazados por cualesquiera otros números. Él explicaría su ejemplo ilustrativo

utilizando símbolos, pero no manipularía los símbolos como tales. Al-Kwārizmī hacía explícita la receta general. Lo hacía utilizando palabras, no símbolos, pero tenía la idea básica, y generalmente es considerado como el padre del álgebra. De hecho, el nombre procede del título de su libro: *Al-Kitāb al-Mukhtasar fī Hisāb al-Jabr wa'Muqābala* («Compendio sobre cálculo por compleción y comparación»). *Al-Jabr* se transformó en «álgebra». La palabra «algoritmo» procede de una versión medieval de su nombre, Algorismus, y ahora se utiliza con el significado de un proceso matemático específico para resolver un problema, un proceso que garantiza que se encontrará una solución siempre que se espere el tiempo suficiente.

Tradicionalmente los matemáticos consideraban que un problema estaba resuelto si, en principio, podían escribir un algoritmo que llevara a una respuesta. Apenas utilizaban esa palabra, y preferían presentar, digamos, una fórmula para la solución, que es un tipo particular de algoritmo en lenguaje simbólico. Que fuera posible aplicar la fórmula en la práctica no era muy importante: la fórmula era la solución. Pero el uso de ordenadores cambió esa visión, porque fórmulas que habían sido demasiado complicadas para calcularse a mano podían hacerse prácticas con la ayuda de un ordenador. Sin embargo, era un poco desagradable encontrar, como sucedía a veces, que la fórmula seguía siendo demasiado complicada: aunque el ordenador podía tratar de ejecutar el algoritmo, era demasiado lento para alcanzar la respuesta. Por ello la atención se desplazó a encontrar algoritmos eficientes. Tanto

matemáticos como científicos de la computación tenían un especial interés en desarrollar algoritmos que realmente dieran respuestas en un período de tiempo razonable.

Dado un algoritmo, es relativamente sencillo calcular cuánto tiempo necesitará (medido por el número de pasos computacionales necesarios) para resolver un problema con una entrada de tamaño dado. Eso puede requerir cierta dosis de técnica, pero uno sabe qué proceso está implicado y sabe mucho sobre lo que se está haciendo. Mucho más difícil es concebir un algoritmo más eficiente si el algoritmo del que se parte resulta ser ineficiente. Y todavía más difícil es decidir cuán bueno o malo pueda ser el algoritmo más eficiente en el caso de un problema dado, porque eso supone contemplar todos los algoritmos posibles, y no se sabe cuáles son.

El trabajo inicial sobre tales cuestiones llevó a una dicotomía tosca pero conveniente entre algoritmos que eran ineficientes, en un sentido simple aunque primario, y los que no lo eran. Si la longitud de la computación crece de forma relativamente lenta cuando aumenta el tamaño del *input*, el algoritmo es eficiente y el problema es fácil. Si la longitud de la computación crece cada vez más deprisa cuando aumenta el tamaño del *input*, el algoritmo es ineficiente y el problema es difícil. La experiencia nos dice que aunque algunos problemas son fáciles en este sentido, la mayoría parecen ser difíciles. De hecho, si todos los problemas matemáticos fueran fáciles, los matemáticos perderían su empleo. El problema del premio del milenio pide una demostración rigurosa de que existe al menos un problema difícil; o de que, contrariamente a la

experiencia, todos los problemas son fáciles. Se conoce como el problema P/NP, y nadie tiene una clave acerca de cómo resolverlo.

Ya hemos encontramos una medida muy tosca de la eficiencia en el capítulo 2. Un algoritmo es de clase P si tiene un tiempo de ejecución polinómico. En otras palabras, el número de pasos que necesita para llegar a la respuesta es proporcional a alguna potencia fija, tal como el cuadrado o el cubo, del tamaño de los datos de entrada. Tales algoritmos son eficientes, hablando en términos muy generales. Si el *input* es un número, dicho tamaño es cuántos dígitos tiene, no el número propiamente dicho. La razón es que la cantidad de información necesaria para especificar el número es el espacio que ocupa en la memoria del ordenador, que es (proporcional a) el número de dígitos. Un problema es de clase P si existe un algoritmo de clase P que lo resuelve.

Cualquier otro algoritmo o problema pertenece a la clase no-P, y la mayoría de estos son ineficientes. Entre ellos están aquellos para los que el tiempo de ejecución es exponencial en los datos de entrada: aproximadamente igual a cierto número fijo elevado a la potencia del tamaño de la entrada. Estos son de clase E, y son decididamente ineficientes.

Algunos algoritmos son tan eficientes que necesitan un tiempo de ejecución mucho menor que polinómico. Por ejemplo, para determinar si un número es par o impar examinamos su último dígito. Si (en notación decimal) este es 0, 2, 4, 6 u 8, el número es par; de lo contrario es impar. El algoritmo consta como máximo de seis pasos:

1. ¿Es el último dígito 0? Si sí, entonces *STOP*. El número es par.
2. ¿Es el último dígito 2? Si sí, entonces *STOP*. El número es par.
3. ¿Es el último dígito 4? Si sí, entonces *STOP*. El número es par.
4. ¿Es el último dígito 6? Si sí, entonces *STOP*. El número es par.
5. ¿Es el último dígito 8? Si sí, entonces *STOP*. El número es par.
6. *STOP*. El número es impar.

Así que el tiempo de ejecución es como máximo 6, independientemente del tamaño del *input*. Pertenece a la clase «tiempo constante».

Ordenar una lista de palabras en orden alfabético es un problema de clase P. Una forma sencilla de realizar esta tarea es el ordenamiento de burbuja, así llamado porque las palabras ascienden en la lista como burbujas en un vaso de bebida gaseosa si están en la lista por debajo de otras palabras que deberían ir detrás de ellas en orden alfabético. El algoritmo recorre repetidamente la lista, compara palabras adyacentes y las intercambia si están en el orden incorrecto. Por ejemplo, supongamos que la lista empieza como

PIG DOG CAT APE

En el primer recorrido esto se convierte en

DOG PIG CAT APE
DOG **CAT PIG** APE
DOG CAT **APE PIG**

donde las palabras en negrita son las que acaban de ser comparadas.
En la segunda ejecución esto se convierte en

CAT DOG APE PIG
CAT **APE DOG** PIG
CAT APE **DOG PIG**

La tercera ejecución da

APE CAT DOG PIG
APE **CAT DOG** PIG
APE CAT **DOG PIG**

En la cuarta ejecución nada cambia y por eso sabemos que hemos terminado. Nótese cómo APE sube como una burbuja paso a paso hasta la parte superior (es decir, al frente).
Con cuatro palabras, el algoritmo recorre tres comparaciones en cada etapa, y hay cuatro etapas. Con *n* palabras, hay *n* - 1

comparaciones por etapa y n etapas, un total de $n(n - 1)$ pasos. Esto es algo menos que n^2, de modo que el tiempo de ejecución es polinómico, de hecho cuadrático. El algoritmo puede acabar antes, pero en el peor de los casos, cuando las palabras están exactamente en orden inverso, necesita $n(n - 1)$ pasos. El ordenamiento de burbuja es obvio y de clase P, pero en modo alguno está cerca del algoritmo de ordenación más eficiente. La ordenación por comparación más rápida, que se establece de una manera más inteligente, se ejecuta en $n \log n$ pasos.

Un algoritmo simple con tiempo de ejecución exponencial, de clase E, es «imprimir una lista de todos los números binarios con n dígitos». Hay 2^n números en la lista, e imprimir cada uno de ellos (y calcularlos) necesita aproximadamente n pasos, de modo que el tiempo de ejecución es aproximadamente $2n^n$, que es mayor que 2^n pero menor que 3^n cuando n es suficientemente grande. Sin embargo, este ejemplo es un poco tonto porque lo que lo hace tan lento es el tamaño de la salida, no la complejidad del cálculo, y esta observación resultará crucial más adelante.

Un algoritmo de clase E más típico resuelve el problema del viajante. Un viajante tiene que visitar varias ciudades. Puede hacerlo en cualquier orden. ¿Qué ruta pasa por todas ellas recorriendo la distancia total más corta? La manera ingenua de resolverlo es hacer una lista de todas las rutas posibles, calcular la distancia total en cada una y encontrar la más corta. Con n ciudades hay

$$n! = n\times(n - 1)\times(n - 2)\times\ldots\times3\times2\times1$$

rutas (léase «*n* factorial»). Esto crece más rápidamente que cualquier exponencial[74]. Un método más eficiente, llamado programación dinámica, resuelve el problema del viajante en tiempo exponencial. El primero de estos métodos, al algoritmo de Held-Karp, encuentra el recorrido más corto en $2^n\ n^2$ pasos, que de nuevo está entre 2^n y 3^n cuando n es suficientemente grande.

Pese a que estos algoritmos son «ineficientes», pueden utilizarse trucos especiales para acortar la computación cuando el número de ciudades es grande para los estándares humanos, pero no demasiado grande para que los trucos dejen de ser efectivos. En 2006 D. L. Applegate, R. M. Bixby, V. Chvátal y W. J. Cook resolvieron el problema del viajante para 85.900 ciudades, y este todavía era el récord a mediados de 2012[75].

Estos ejemplos de algoritmos no solo ilustran el concepto de eficiencia. También transmiten la idea que quiero exponer sobre la dificultad de encontrar uno que sea lo más eficiente posible. Todos los algoritmos conocidos para el problema del viajante son de clase E, tiempo exponencial, pero eso no implica que no exista ningún algoritmo eficiente. Solo muestra que todavía no hemos encontrado uno. Hay dos posibilidades: no hemos encontrado un algoritmo

[74] La fórmula de Stirling afirma que *n*! es aproximadamente

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

[75] William J. Cook, *In Pursuit of the Travelling Salesman*, Princeton University Press, Princeton, 2012. Para información actual, véase http:// www.tsp.gatech.edu/index.html.

mejor porque no somos lo bastante inteligentes, o no hemos encontrado un algoritmo mejor porque no existe.

El capítulo 2 hace al caso. Hasta que el equipo de Agrawal encontró su algoritmo de clase P para comprobar la primalidad, el algoritmo más conocido era no P. Era aún bastante bueno, con un tiempo de ejecución $n \log n$ para números de n dígitos, que realmente es mejor que el algoritmo de Agrawal-Kayal-Saxena hasta que llegamos a números con diez mil dígitos. Antes de que se descubriera su algoritmo, la opinión sobre el estatus de la comprobación de primalidad estaba dividida. Algunos expertos sospechaban que era de clase P y se encontraría un algoritmo adecuado; otros pensaban que no lo era. El nuevo algoritmo salió de la nada, una de los miles de ideas que alguien podría haber ensayado; resultó que esta funcionaba. Aquí el precedente no es muy alentador: no sabemos, no podemos distinguir, y la mejor conjetura de los expertos puede ser buena o no serlo.

El gran problema que aquí nos interesa busca la respuesta a una pregunta más fundamental. ¿Existen problemas difíciles? ¿Podrían ser fáciles todos los problemas si fuéramos suficientemente inteligentes? El enunciado real es más sutil, porque ya hemos visto un caso de un problema que es sin lugar a dudas difícil: imprimir una lista de todos los números binarios con n dígitos. Como he comentado, esto es un poco tonto: la dificultad no reside en el cálculo, sino en la tarea tremendamente aburrida de imprimir una respuesta muy larga. Sabemos que no hay atajo porque la respuesta

tiene esa longitud *por definición*. Si fuera más corta, no sería la respuesta.

Para plantear una pregunta razonable, ejemplos triviales como este deben ser eliminados. La manera de hacerlo es introducir otra clase de algoritmo, la clase NP. Esta no es la clase no-P; es la clase de algoritmos que se ejecutan en tiempo polinómico no determinista. La jerga significa que, por mucho tiempo que tarde el algoritmo en dar con su respuesta, podemos *comprobar que la respuesta es correcta* en tiempo polinómico. Encontrar la respuesta puede ser difícil, pero, una vez encontrada, hay una comprobación fácil de su validez.

La expresión «no determinista» se utiliza aquí porque es posible resolver un problema NP haciendo una conjetura inspirada. Una vez hecha, es posible confirmar si realmente es correcta (o que no lo es). Por ejemplo, si el problema es factorizar el número 11.111.111.111, se podría conjeturar que un factor es el número primo 21.649. Tal como está, eso es tan solo una conjetura aventurada. Pero es fácil de comprobar: simplemente dividimos por dicho número y vemos lo que se obtiene. El resultado es 513.239, exactamente, sin resto. De modo que la conjetura era correcta. Si en su lugar hubiéramos conjeturado 21.647, que también es primo, entonces la división llevaría al resultado 513.286 más un resto de 9069. De modo que dicha conjetura habría sido errónea.

Hacer una conjetura correcta aquí es básicamente un milagro, o de lo contrario hay un truco (yo calculé los factores de 11.111.111.111 antes de «conjeturar»). Pero eso es, de hecho, lo que buscamos. Si

no fuera milagroso, se podría convertir un algoritmo de clase NP en un algoritmo de clase P con solo hacer montones y montones de conjeturas hasta que una resulte ser correcta. Mi ejemplo sugiere por qué esto no funciona: hay que hacer demasiadas conjeturas. De hecho, todo lo que estamos haciendo aquí es «ensayar la división» por todos los primos posibles hasta que uno funcione. Sabemos del capítulo 2 que esto es una manera desesperada de encontrar factores.

La clase NP descarta ejemplos estúpidos como mi larga lista. Si alguien conjetura una lista de todos los dígitos binarios de longitud *n*, entonces no solo se necesita un tiempo exponencial para imprimir la lista. También se necesita un tiempo exponencial para leerla, de modo que se tarda un tiempo aún mayor en comprobar si es correcta. Sería una tarea realmente horrible de lectura de pruebas. La clase P está definitivamente contenida en la clase NP. Si se puede encontrar la respuesta en tiempo polinómico, con una garantía de que es correcta, entonces ya se ha comprobado. De modo que la comprobación no requiere nada peor que tiempo polinómico. Si alguien nos presentara la respuesta supuesta, podríamos simplemente ejecutar de nuevo el algoritmo completo. Esa es la comprobación.

Ahora podemos enunciar el problema del milenio. ¿Es NP más grande que P, o son iguales? Más brevemente: ¿es P igual a NP?

Si la respuesta es «sí», entonces sería posible encontrar algoritmos rápidos y eficientes para programar los vuelos de las líneas aéreas, optimizar la producción de las fábricas o realizar millones de otras

importantes tareas prácticas. Si la respuesta es «no», tendremos una garantía férrea de que todos los problemas que parecen difíciles son realmente difíciles, y así podremos dejar de perder tiempo tratando de encontrar algoritmos rápidos para ellos. En uno u otro caso, algo ganamos. Lo que es una molestia es no saber cuál es el caso.

La vida de los matemáticos sería mucho más sencilla si la respuesta fuera «sí», pero el pesimista que hay en todo ser humano sospecha que la vida no va a ser tan sencilla, y que probablemente la respuesta sea «no». De lo contrario todos tenemos barra libre, lo que no merecemos ni nos hemos ganado. Sospecho que la mayoría de los matemáticos preferirían que la respuesta sea «no», porque eso les mantendría ocupados hasta el fin de la civilización. Los matemáticos se prueban a sí mismos resolviendo problemas difíciles. Sea cual sea la razón, la mayoría de los matemáticos y científicos de la computación esperan que la respuesta a la pregunta «¿es P igual a NP?» sea «no». Casi nadie espera que sea «sí».

Existen otras dos posibilidades. Quizá fuera posible demostrar que P es igual a NP sin encontrar realmente un algoritmo de tiempo polinómico para cualquier problema NP concreto. Las matemáticas tienen costumbre de ofrecer demostraciones de existencia que no son constructivas; demuestran que algo existe pero no nos dicen cómo es. Ejemplos de ello son los testes de primalidad que nos informan alegremente de que un número no es primo sin darnos ningún factor concreto, o teoremas de la teoría de números que aseveran que las soluciones de cierta ecuación diofántica son acotadas —menores que un cierto valor límite— sin dar ninguna

cota concreta. Un algoritmo de tiempo polinómico podría ser tan complicado que fuera imposible desarrollarlo. Entonces el pesimismo natural sobre las barras libres estaría justificado incluso si la respuesta resultara ser afirmativa.

De manera más drástica, algunos investigadores especulan con que la pregunta pueda ser indecidible dentro del marco actual de la lógica formal para las matemáticas. Si es así, ni «sí» ni «no» pueden demostrarse. No porque seamos demasiado estúpidos para encontrar la demostración, sino porque no la hay. Esta opción se hizo patente en 1931 cuando Kurt Gödel dejó suelto el gato de la indecibilidad entre las palomas filosóficas que infestaban los fundamentos de las matemáticas, al demostrar que algunas proposiciones en aritmética son indecidibles. En 1936 Alan Turing encontró un problema indecidible más simple, el problema de la parada para máquinas de Turing. Dado un algoritmo, ¿hay siempre una demostración de que se para, o una demostración de que debe continuar indefinidamente? La sorprendente respuesta de Turing fue «no». Para algunos algoritmos no existe demostración ni en un sentido ni en el otro. Tal vez el problema P/NP pudiera ser así. Eso explicaría por qué nadie puede demostrarlo ni refutarlo. Pero tampoco nadie puede demostrar ni refutar que el problema P/NP sea indecidible. Quizá su indecibilidad sea indecidible…

La forma más directa de acercarse al problema P/NP sería seleccionar alguna pregunta que se sabe que es de clase NP, suponer que existe un algoritmo de tiempo polinómico para resolverla, y de algún modo derivar una contradicción. Durante un

tiempo se ensayó está técnica con varios problemas, pero en 1971 Stephen Cook se dio cuenta de que la elección del problema no suele suponer ninguna diferencia. Hay un sentido en el que todos estos problemas —con más o menos tecnicismos— se resisten o caen juntos. Cook introdujo la noción de un problema NP completo. Este es un problema NP específico con la propiedad de que si existe un algoritmo de clase P para resolverlo, entonces *cualquier* problema NP puede resolverse utilizando un algoritmo de clase P.

Cook encontró varios problemas NP completos, incluido SAT, el problema de la satisfacibilidad booleana. Este pregunta si una expresión lógica dada puede hacerse verdadera escogiendo la verdad o falsedad de sus variables de una manera apropiada. También obtuvo un resultado más profundo: un problema más restrictivo, 3-SAT, es también NP-completo. Aquí la fórmula lógica es una que puede escribirse de la forma «*A* o *B* o *C* o … o *Z*», donde cada uno de los *A, B, C*…, *Z* es una fórmula lógica que implica solo tres variables. No necesariamente los tres mismas variables cada vez, me apresuro a añadir. La mayoría de las demostraciones de que un problema dado es NP-completo se remontan al teorema de Cook sobre 3-SAT.

La definición de Cook implica que todos los problemas NP-completos están en pie de igualdad. Demostrar que uno de ellos es de clase P demostraría que todos ellos son de clase P. Este resultado deja abierta una posibilidad táctica: algunos problemas NP-completos podrían ser más fáciles de tratar que otros. Pero desde el punto de vista estratégico sugiere que también se puede escoger un problema NP-completo y trabajar con él. Todos los problemas NP-completos se

resisten o caen juntos porque un problema NP-completo puede simular cualquier problema NP. Cualquier problema NP puede convertirse en un caso especial de problema NP-completo «codificándolo», utilizando un código que puede implementarse en tiempo polinómico.

Para dar una idea de este procedimiento, consideremos un típico problema NP-completo: encontrar un ciclo hamiltoniano en una red. Es decir, especifiquemos un camino cerrado a lo largo de las aristas de la red que visita cada vértice (punto) exactamente una vez. Cerrado significa que el camino vuelve a su punto de partida. El tamaño de los datos de entrada es aquí el número de aristas, que es menor o igual que el cuadrado del número de puntos puesto que cada arista une dos puntos. (Suponemos que como mucho hay una arista uniendo un par dado). No se conoce ningún algoritmo de clase P para resolver este problema, pero supongamos, hipotéticamente, que hubiera uno. Ahora escojamos algún otro problema y llamémosle problema X. Supongamos que el problema X puede reformularse en términos de encontrar un camino semejante en una red asociada al problema X. Si el método para traducir los datos del problema X en datos sobre la red, y recíprocamente, puede realizarse en tiempo polinómico, entonces obtenemos automáticamente un algoritmo de clase P para el problema X, como este:

1. Traducir el problema X en la búsqueda de un ciclo hamiltoniano en la red relacionada, lo que puede hacerse en tiempo polinómico.

2. Encontrar dicho ciclo en tiempo polinómico utilizando el algoritmo hipotético para el problema de la red.
3. Traducir el ciclo hamiltoniano resultante en una solución del problema X, lo que de nuevo puede hacerse en tiempo polinómico.

Puesto que tres pasos de tiempo polinómico combinados se ejecutan en tiempo polinómico, este algoritmo es de clase P.

Para mostrar cómo trabaja esto voy a considerar una versión menos ambiciosa del problema del ciclo hamiltoniano en la que no se exige que el camino sea cerrado. Esto se denomina el problema del camino hamiltoniano. Una red puede poseer un camino hamiltoniano sin poseer un ciclo: la Figura 42 (izquierda) es un ejemplo. Por ello es posible que una solución del problema del ciclo hamiltoniano no resuelva el problema del camino hamiltoniano. Sin embargo, podemos convertir el problema del camino hamiltoniano en un problema de ciclo hamiltoniano en una red relacionada aunque diferente. Esto se consigue añadiendo un punto extra, unido a cada punto en la red original como en la Figura 42 (derecha). Cualquier ciclo hamiltoniano en la nueva red puede convertirse en un camino hamiltoniano en la red original: basta con omitir el nuevo vértice y las dos aristas del ciclo que se encuentran en él. Recíprocamente, cualquier camino hamiltoniano en la red original da un ciclo hamiltoniano en la nueva red: basta con unir los dos extremos del camino hamiltoniano al nuevo punto. Esta «codificación» del problema del camino como un problema de ciclo introduce solo un punto nuevo y una arista nueva por punto en el

original. Por lo tanto, este procedimiento, y su inverso, se ejecutan en tiempo polinómico.

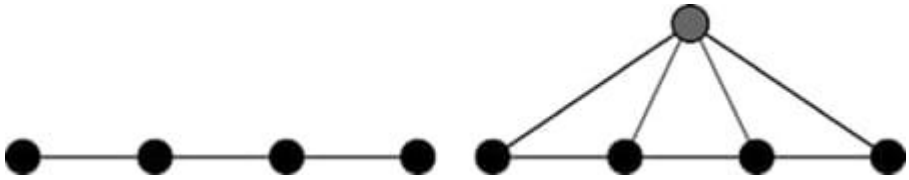


Figura 42. Izquierda: Red con un camino hamiltoniano (línea sólida) pero sin ciclo hamiltoniano. Derecha: Añadir un punto extra (gris) y cuatro líneas más para convertir el camino hamiltoniano en un ciclo hamiltoniano (línea sólida). Las dos aristas grises no están en el ciclo pero son necesarias para la construcción de la red más grande.

Por supuesto, todo lo que he hecho aquí es codificar un problema específico como un problema de ciclo hamiltoniano. Para demostrar que el problema del ciclo hamiltoniano es NP-completo tenemos que hacer lo mismo para cualquier problema NP. Esto puede hacerse: la primera demostración fue encontrada por Richard Karp en 1972, en un famoso artículo que demostraba que 21 problemas diferentes eran NP-completos[76].

El problema del viajante es «casi» NP-completo, pero hay una cuestión técnica: no se sabe que sea NP. Se conocen más de trescientos problemas NP-completos en áreas de las matemáticas que incluyen la lógica, las redes, la combinatoria y la optimización. Demostrar que cualquiera de ellos puede o no puede resolverse en tiempo polinómico demostraría lo mismo para cualquiera de ellos.

[76] Richard M. Karp, «Reducibility among combinatorial problems», en R. F. Miller y J. W. Thatcher (eds.), *Complexity of Computer Computations*, Plenum, 1972, pp. 8-103.

Pese a que hay mucho donde elegir, el problema P/NP sigue totalmente abierto. No sería para mí una sorpresa que lo siga estando dentro de cien años.

Capítulo 12

Pensamiento fluido

La ecuación de Navier-Stokes

Cinco de los problemas del milenio, incluidos los tres discutidos hasta ahora, proceden de las matemáticas puras, aunque el problema P/NP es también fundamental para las ciencias de la computación. Los otros dos proceden de las matemáticas aplicadas clásicas y de la moderna física matemática. El problema de las matemáticas aplicadas surge de una ecuación estándar para el flujo de un fluido, la ecuación de Navier-Stokes, que debe su nombre al ingeniero y físico francés Claude-Louis Navier y el matemático y físico irlandés Georges Stokes. Su ecuación es una ecuación en derivadas parciales, lo que significa que incluye el ritmo de cambio del flujo en el espacio y en el tiempo. La mayoría de las grandes ecuaciones de las matemáticas aplicadas y de la física son también ecuaciones en derivadas parciales; acabamos de encontrar una, la de Laplace. Y las que no son de este tipo, son ecuaciones diferenciales ordinarias, que solo implican el ritmo de cambio con respecto al tiempo.

En el capítulo 8 vimos cómo el movimiento del Sistema Solar está determinado por las leyes de Newton de la gravedad y el movimiento. Estas relacionan las aceleraciones del Sol, la Luna y los planetas con las fuerzas gravitatorias que están actuando. La aceleración es el ritmo de cambio de la velocidad con respecto al tiempo, y la velocidad es el ritmo de cambio de la posición con

respecto al tiempo. Por lo tanto, esta es una ecuación diferencial ordinaria. Como vimos, resolver tales ecuaciones puede ser muy difícil. Resolver ecuaciones en derivadas parciales es en general mucho más difícil.

Para fines prácticos, las ecuaciones para el Sistema Solar pueden resolverse numéricamente utilizando ordenadores. Esto sigue siendo difícil, pero ahora existen buenos métodos. Lo mismo es cierto para las aplicaciones prácticas de las ecuaciones de Navier-Stokes. Las técnicas utilizadas se conocen como dinámica de fluidos computacional, y tienen un amplio espectro de importantes aplicaciones: diseño de aviones, aerodinámica de automóviles, incluso problemas médicos como el flujo sanguíneo en el cuerpo humano.

El problema del premio del milenio no pide a los matemáticos que encuentren soluciones explícitas a la ecuación de Navier-Stokes, ya que esto es esencialmente imposible. Ni trata de métodos numéricos para resolver las ecuaciones, por importantes que sean estos. Lo que pide, en su lugar, es una demostración de una propiedad teórica básica: la *existencia* de soluciones. Dado el estado de un fluido en algún instante de tiempo —su patrón de movimiento—, ¿existe una solución de la ecuación de Navier-Stokes válida para cualquier instante futuro, a partir del estado en cuestión? La intuición física sugiere que la respuesta debe ser sin duda «sí», porque la ecuación es un modelo muy preciso de la física de los fluidos reales. Sin embargo, la cuestión matemática de la existencia

no está tan clara, y esta propiedad básica de la ecuación nunca ha sido demostrada. Incluso podría no ser cierta.

La ecuación de Navier-Stokes describe cómo cambia con el tiempo la pauta de velocidades del fluido, en circunstancias dadas. A menudo se hace referencia a la ecuación utilizando el plural, ecuaciones de Navier-Stokes, pero es lo mismo. El plural refleja la visión clásica: en el espacio tridimensional la velocidad tiene tres componentes, y clásicamente cada componente aporta una ecuación, lo que hace tres en total. En la visión moderna hay una ecuación para el *vector* velocidad (una magnitud con tamaño y dirección), pero esta ecuación puede aplicarse a cada una de las tres componentes de la velocidad. La página web del Instituto Clay utiliza la terminología clásica, pero yo seguiré aquí la práctica moderna. Lo menciono para evitar posibles confusiones.

La ecuación data de 1822, cuando Navier escribió una ecuación en derivadas parciales para el flujo de un fluido viscoso (pegajoso). Las contribuciones de Stokes son de 1842 y 1843. Euler había escrito una ecuación en derivadas parciales para un fluido con viscosidad nula (no pegajoso) en 1757. Aunque esta ecuación sigue siendo útil, la mayoría de los fluidos reales, agua y aire incluidos, son viscosos, de modo que Navier y Stokes modificaron la ecuación de Euler para tener en cuenta la viscosidad. Ambos científicos derivaron, de forma independiente, esencialmente la misma ecuación, por lo que esta lleva el nombre de ambos. Navier cometió algunos errores matemáticos, pero terminó con la respuesta correcta; Stokes hizo bien las matemáticas, y por eso sabemos que la respuesta de Navier

es correcta pese a su error. En su forma más general, la ecuación se aplica a fluidos compresibles como el aire. Sin embargo, hay un caso especial importante en el que se supone que el fluido es incompresible. Este modelo se aplica a fluidos como el agua, que se comprime cuando se somete a fuerzas enormes, aunque muy ligeramente.

Hay dos maneras de describir matemáticamente el flujo de un fluido: se puede describir la trayectoria que sigue cada partícula del fluido con el paso del tiempo, o se puede describir la velocidad del flujo en cada punto del espacio y en cada instante de tiempo. Las dos descripciones están relacionadas: dada una, se puede —con algún esfuerzo— deducir la otra. Euler, Navier y Stokes utilizaron el segundo punto de vista porque lleva a una ecuación que es mucho más tratable matemáticamente. Por eso sus ecuaciones se refieren al campo de velocidades del fluido. En cada instante de tiempo, el campo de velocidades especifica la velocidad y dirección de cada partícula del fluido. Conforme pasa el tiempo, esta descripción puede cambiar. Por esto es por lo que en la ecuación aparecen los ritmos de cambio en el espacio y en el tiempo.

La ecuación de Navier-Stokes tiene un excelente pedigrí físico. Se basa en las leyes de movimiento de Newton aplicadas a cada minúscula partícula (pequeña región) de fluido, y expresa, en dicho contexto, la ley de conservación del momento lineal. Cada partícula se mueve porque sobre ella actúan fuerzas y la ley de movimiento de Newton afirma que la aceleración de la partícula es proporcional a la fuerza. Las fuerzas principales son la fricción, debida a la

viscosidad, y la presión. También hay fuerzas generadas por la aceleración de la partícula. La ecuación sigue la práctica clásica y trata el fluido como un continuo infinitamente divisible. En particular, ignora la estructura atómica discreta del fluido en escalas muy pequeñas.

Las ecuaciones por sí solas son de poco valor: hay que ser capaces de resolverlas. En el caso de la ecuación de Navier-Stokes esto significa calcular el campo de velocidades: la velocidad y dirección del fluido en cada punto del espacio y en cada instante en el tiempo. La ecuación proporciona ligaduras sobre estas cantidades, pero no las prescribe directamente. En su lugar, tenemos que aplicar la ecuación para relacionar las velocidades futuras con las actuales. Las ecuaciones en derivadas parciales como las de Navier-Stokes tienen muchas soluciones diferentes; de hecho, infinitas soluciones. Esto no es una sorpresa: los fluidos pueden fluir de muchas maneras diferentes; el flujo sobre la superficie de un automóvil difiere del flujo sobre las alas de un avión. Hay dos maneras principales de seleccionar un flujo particular de entre esta multitud de posibilidades: por las condiciones iniciales y por las condiciones de contorno.

Las condiciones iniciales especifican el campo de velocidades en un tiempo de referencia particular, que normalmente se toma como tiempo cero. La idea física es que una vez que se sabe el campo de velocidades en ese instante, la ecuación de Navier-Stokes determina de modo unívoco el campo muy poco tiempo después. Si se empieza dando al fluido un empujón, sigue en marcha aunque obedeciendo

las leyes de la física. Las condiciones de contorno son más útiles en la mayoría de las aplicaciones, porque es difícil establecer condiciones iniciales en un fluido real, y en cualquier caso estas no son por completo apropiadas para aplicaciones como el diseño de automóviles. Lo que importa aquí es la forma del automóvil. El fluido viscoso se adhiere a las superficies. Matemáticamente, esta propiedad se modela especificando la velocidad en estas superficies, que forman el contorno de la región ocupada por el fluido donde la ecuación es válida. Por ejemplo, podríamos exigir que la velocidad sea cero en el contorno, o cualquier otra condición que mejor modele la realidad.

Incluso cuando se especifican condiciones iniciales o condiciones de contorno, es muy inusual poder escribir una fórmula explícita para el campo de velocidades, porque la ecuación de Navier-Stokes es no lineal. La suma de dos soluciones no es normalmente una solución. Esta es una razón por la que el problema de los tres cuerpos del capítulo 8 es tan difícil, aunque no la única razón pues el problema de dos cuerpos también es no lineal pero tiene una solución explícita.

Para fines prácticos podemos resolver la ecuación de NavierStokes en un ordenador y representar el campo de velocidades como una lista de números. Esta lista puede convertirse en gráficos elegantes y utilizarse para calcular cantidades de interés para los ingenieros, tales como las tensiones en las alas de un avión. Puesto que los ordenadores no pueden procesar listas infinitas de números, ni pueden procesar números con precisión infinita, tenemos que

reemplazar el flujo real por una aproximación discreta, es decir, una lista de números que es una muestra del flujo en un número finito de lugares e instantes. La gran cuestión está en asegurar que la aproximación es suficientemente buena.

El enfoque usual consiste en dividir el espacio en un gran número de regiones pequeñas para formar una malla computacional. La velocidad se calcula solo para los puntos en los nodos de la malla. La malla podría ser simplemente un conjunto de cuadrados (o de cubos en tres dimensiones), como un tablero de ajedrez, pero en el caso de automóviles y aviones tiene que ser más complicada, con regiones más pequeñas cerca del contorno para captar los detalles más finos del flujo. La malla puede ser dinámica, cambiando de forma con el paso del tiempo. En general se supone que el tiempo avanza en pasos, que pueden ser todos del mismo tamaño o pueden cambiar de tamaño según el estado imperante del cálculo.

La base de la mayoría de los métodos numéricos es la forma en que se define «ritmo de cambio» en el cálculo infinitesimal. Supongamos que un objeto se mueve de un lugar a otro en un período de tiempo muy corto. Entonces el ritmo de cambio de la posición —la velocidad— es el cambio en la posición dividido por el tiempo que ha tardado, con un pequeño error que disminuye a medida que el intervalo de tiempo se hace cada vez menor. Así que podemos aproximar el ritmo de cambio, que es lo que entra en la ecuación de Navier-Stokes, por esta razón entre el cambio espacial y el cambio temporal. En efecto, la ecuación nos dice ahora cómo llevar un estado inicial conocido —una lista de velocidades especificada— un

paso de tiempo hacia el futuro. Luego tenemos que repetir el cálculo muchas veces para ver qué sucede en un futuro cada vez más lejano. Hay una manera similar de aproximar soluciones cuando la que buscamos está determinada por condiciones de contorno. También existen muchas maneras sofisticadas de llegar al mismo resultado con más precisión.

Cuanto más fina es la malla computacional, y más cortos son los intervalos de tiempo, más precisa se hace la aproximación. Sin embargo, la computación también necesita más tiempo. Por ello hay un compromiso entre precisión y velocidad. Hablando en general, es probable que una respuesta aproximada obtenida por computación sea aceptable siempre que el flujo no tenga características importantes que sean menores que el tamaño de la malla. Existen dos tipos principales de flujo fluido: laminar y turbulento. En el flujo laminar, la pauta del movimiento es suave y las capas de fluido deslizan limpiamente unas al lado de otras. Aquí debería ser apropiada una malla pequeña. El flujo turbulento es más violento y espumoso, y el fluido se mezcla de maneras muy complejas. En tales circunstancias, una malla discreta, por fina que sea, podría causar problemas.

Una de las características de la turbulencia es la aparición de vórtices, como pequeños torbellinos, y estos pueden ser realmente minúsculos. Una imagen estándar de la turbulencia consiste en una cascada de vórtices cada vez más pequeños. La mayor parte del detalle fino es más pequeño que cualquier malla práctica. Para evitar esta dificultad los ingenieros suelen recurrir a modelos

estadísticos cuando se trata del flujo turbulento. Otra preocupación es que el modelo físico de un continuo podría ser inadecuado para el flujo turbulento, porque los vórtices pueden contraerse hasta tamaños atómicos. Sin embargo, comparaciones entre cálculos numéricos y experimentos muestran que la ecuación de Navier-Stokes es un modelo muy realista y aproximado; un modelo tan bueno que muchas aplicaciones en ingeniería actuales se basan solamente en dinámica de fluidos computacional, que es barata, más que en experimentos con modelos a escala en túneles de viento, que son caros. Sin embargo, las comprobaciones experimentales como estas siguen siendo utilizadas cuando la seguridad humana es vital, por ejemplo en el diseño de aviones.

De hecho, la ecuación de Navier-Stokes es tan precisa que incluso parece aplicarse cuando la física sugiere que debería haber una probabilidad razonable de fallar: en el flujo turbulento. Al menos, este es el caso si puede resolverse con suficiente precisión. El problema principal es un problema práctico: los métodos numéricos para resolver la ecuación necesitan mucho tiempo de computación cuando el flujo se hace turbulento. Y siempre dejan fuera alguna estructura a pequeña escala.

Los matemáticos se sienten siempre incómodos cuando la información principal de que disponen para tratar un problema se basa en algún tipo de aproximación. El premio del milenio para la ecuación de Navier-Stokes aborda una de las cuestiones teóricas clave. Su solución reforzaría la sensación visceral de que normalmente los métodos numéricos funcionan muy bien. Hay una

sutil distinción entre las aproximaciones utilizadas por el ordenador, que hacen pequeños cambios en la ecuación, y la exactitud de la respuesta, que se refiere a pequeños cambios en la solución. ¿Es una respuesta exacta a una pregunta aproximada lo mismo que una respuesta aproximada a una pregunta exacta? A veces la respuesta es «no». Por ejemplo, el flujo exacto para un fluido con una viscosidad muy pequeña suele diferir de un flujo aproximado para un fluido con viscosidad cero.

Un paso hacia la comprensión de estas cuestiones es tan sencillo que fácilmente puede pasarse por alto: demostrar que existe una solución exacta. Tiene que haber algo a lo que los cálculos por ordenador son aproximaciones. Esta observación motiva el premio del milenio para la ecuación de Navier-Stokes. Su descripción oficial en la página web del Instituto Clay consiste en cuatro problemas. Resolver cualquiera de ellos es suficiente para ganar el premio. En los cuatro, el fluido se supone incompresible. Son:

1. *Existencia y suavidad de soluciones en tres dimensiones*. Aquí se supone que el fluido llena todo el espacio infinito. Dado cualquier campo de velocidades suave, demostrar que una solución suave de la ecuación existe para cualquier instante positivo, coincidente con el campo inicial especificado.
2. *Existencia y suavidad de soluciones en el toro plano tridimensional*. La misma pregunta, pero ahora suponiendo que el espacio es un toro plano —una caja rectangular con caras opuestas identificadas—. Esta versión evita posibles problemas causados por el dominio infinito supuesto en la

primera versión, que no encaja con la realidad y podría provocar mal comportamiento por tontas razones.

3. *Inexistencia de soluciones en tres dimensiones*. Demostrar que (1) es falso. Es decir, encontrar un campo inicial para el que no existe una solución suave para cualquier instante positivo, y demostrar dicha afirmación. 4. *Inexistencia de soluciones en el toro plano tridimensional*. Demostrar que (2) es falso.

Los mismos problemas siguen abiertos para la ecuación de Euler, que es la misma que la ecuación de Navier-Stokes pero supone que no hay viscosidad. Sin embargo, no se ofrece ningún premio para la ecuación de Euler.

La gran dificultad aquí es que el flujo bajo consideración es tridimensional. Hay una ecuación análoga para el fluido que fluye en un plano. Físicamente, esto representa o bien una capa delgada de fluido entre dos placas planas, que se supone que no causan fricción, o una pauta de flujo en tres dimensiones en la que el fluido se mueve exactamente de la misma manera a lo largo de un sistema de planos paralelos. En 1969 la matemática rusa Olga Alexandrovna Ladyzhenskaya demostró que (1) y (2) son verdaderos, mientras que (3) y (4) son falsos, para la ecuación de Navier-Stokes bidimensional y la ecuación de Euler bidimensional.

Puede parecer sorprendente que la demostración sea más difícil para la ecuación de Euler, incluso si dicha solución es más sencilla que la ecuación de Navier-Stokes pues omite los términos que implican viscosidad. La razón es instructiva. La viscosidad «amortigua» el mal comportamiento en la solución, que

potencialmente podría llevar a algún tipo de singularidad que impida que la solución exista en cualquier instante. Si el término con viscosidad está ausente no puede ocurrir tal amortiguamiento, y esto se manifiesta como cuestiones matemáticas en la demostración de existencia.

Ladyzhenskaya hizo otras contribuciones vitales a nuestra comprensión de la ecuación de Navier-Stokes, al demostrar no solo que esas soluciones existen sino también que ciertos esquemas de dinámica de fluidos computacional se aproximan a ellas tanto como queramos.

El premio del milenio se refiere al flujo incompresible porque es bien sabido que los flujos compresibles tienen un mal comportamiento. Las ecuaciones para un avión, por ejemplo, tropiezan con todo tipo de problemas si el avión va a una velocidad superior a la del sonido. Esta es la famosa «barrera del sonido» que preocupaba a los ingenieros que trataban de diseñar aviones de reacción supersónicos, y el problema está relacionado con la compresibilidad del aire. Si un cuerpo se mueve a través de un fluido incompresible, aparta a las partículas del fluido de su camino, como cuando se hace un túnel a través de una caja llena de bolas de cojinete. Si las partículas se amontonan, frenan al cuerpo. Pero en un fluido compresible, donde hay un límite a la velocidad a la que pueden viajar las ondas (la velocidad del sonido) eso no sucede. A velocidades supersónicas, en lugar de ser apartado, el aire se comprime por delante del avión y su densidad aumenta sin límite. El resultado es una onda de choque. Matemáticamente, esto es una

discontinuidad en la presión del aire, que de repente salta de un valor a otro valor diferente a través de la onda de choque. Físicamente, el resultado es un *boom* sónico: un sonoro *bang*. Si no se entiende y se tiene en cuenta, una onda de choque puede dañar al avión, de modo que los ingenieros tenían razones para preocuparse. Pero la velocidad del sonido no es realmente una barrera, sino solo un obstáculo. La presencia de ondas de choque implica que las ecuaciones de Navier-Stokes compresibles no tienen por qué tener soluciones suaves todo el tiempo, ni siquiera en dos dimensiones. Así que la respuesta ya es conocida en este caso, y es negativa.

Las matemáticas de las ondas de choque son un área sustancial dentro de las ecuaciones en derivadas parciales, pese a esta inexistencia de soluciones. Aunque la ecuación de Navier-Stokes por sí sola no es un buen modelo físico para fluidos compresibles, es posible modificar el modelo matemático añadiendo condiciones extra a las ecuaciones que tienen en cuenta discontinuidades de ondas de choque. Pero las ondas de choque no se dan en el flujo de un fluido incompresible, de modo que es al menos concebible que en dicho contexto deberían existir soluciones en todo tiempo, por muy complicado que pudiera ser el flujo inicial, siempre que sea suave.

Se conocen algunos resultados posibles para la ecuación de Navier-Stokes tridimensional. Si la pauta de flujo inicial implica velocidades suficientemente pequeñas, de modo que el flujo es muy lento, entonces (1) y (2) son verdaderos. Incluso si las velocidades

son grandes, (1) y (2) son verdaderos para un intervalo de tiempo no nulo. Puede no existir una solución válida para todo tiempo futuro, pero hay un intervalo de tiempo definido durante el que existe una solución. Podría parecer que podemos repetir este proceso, avanzando una solución en el tiempo en pequeñas cantidades y luego utilizar el resultado final como una nueva condición inicial. El problema con esta línea de razonamiento es que los intervalos de tiempo pueden contraerse tan rápidamente que un número infinito de tales pasos tardan un tiempo finito. Por ejemplo, si cada paso sucesivo tarda la mitad de tiempo del anterior, y el primer paso tarda, digamos, 1 minuto, entonces el proceso total termina en un tiempo 1 + ½ + ¼ + ⅛ + …, que es igual a 2. Si la solución deja de existir —una suposición puramente hipotética por el momento, pero que cabe contemplar— entonces se dice que la solución en cuestión *explota*. El tiempo que tarda en que esto suceda es el tiempo de explosión.

Así que las cuatro cuestiones preguntan si las soluciones pueden explotar. Si no pueden, (1) y (2) son verdaderos; si pueden, (3) y (4) lo son. Quizá las soluciones pueden explotar en un dominio infinito, pero no en uno finito. De paso, si la respuesta a (1) es «sí», entonces también lo es la respuesta a (2), porque podemos interpretar cualquier pauta de flujo en un toro plano como una pauta de flujo espacialmente periódica en la totalidad del espacio infinito. La idea es llenar el espacio con copias de la caja rectangular implicada y copiar la misma pauta de flujo en cada una de ellas. Las reglas de unión para un toro aseguran que el flujo sigue siendo suave cuando

cruza estas interfaces planas. De modo análogo, si la respuesta a (4) es «sí», entonces también lo es la respuesta a (3), por la misma razón. Tan solo debemos hacer el estado inicial espacialmente periódico. Pero por todo lo que sabemos hoy día, la respuesta a (2) podría ser «sí» pero la respuesta a (1) podría ser «no».

Sabemos un hecho sorprendente sobre las explosiones. Si hay una solución con un tiempo de explosión finito, entonces la máxima velocidad del fluido, en todos los puntos del espacio, debe hacerse arbitrariamente grande. Esto podría ocurrir, por ejemplo, si se forma un chorro de fluido y la velocidad del chorro aumenta tan rápidamente que diverge a infinito una vez que haya pasado una cantidad de tiempo finita.

Estas objeciones no son puramente hipotéticas. Hay precedentes de este tipo de comportamiento singular en otras ecuaciones de la física matemática clásica. Un ejemplo notable se da en la mecánica celeste. En 1988 Zhihong Xia demostró que existe una configuración inicial de cinco masas puntuales en el espacio tridimensional, que obedece a la ley de la gravedad de Newton, para la que cuatro partículas desaparecen en el infinito tras un período de tiempo finito —una forma de explosión— y la quinta sufre oscilaciones cada vez más violentas. Previamente, Joseph Gerver había indicado que los cinco cuerpos en un plano podrían desaparecer en el infinito en un tiempo finito, pero no pudo completar la demostración en el escenario que él imaginaba. En 1989 demostró que este tipo de escape podía ocurrir de hecho en un plano si el número de cuerpos es suficientemente grande.

Es notable que este comportamiento sea posible, dado que tales sistemas obedecen la ley de conservación de la energía. ¿No debería aumentar la energía cinética total si todos los cuerpos se mueven con rapidez arbitraria? La respuesta es que también hay una disminución de la energía potencial, y para una partícula puntual la energía potencial gravitatoria total es infinita. Los cuerpos también deben conservar el momento angular, pero pueden hacerlo con tal de que algunos de ellos se muevan cada vez más rápidos en círculos siempre decrecientes.

El punto físico importante aquí implicado es el famoso efecto honda, que se utiliza rutinariamente para lanzar sondas espaciales a mundos distantes en el Sistema Solar. Un buen ejemplo es la sonda *Galileo* de la NASA, cuya misión era viajar a Júpiter para estudiar el planeta gigante y sus muchos satélites. Fue lanzada en 1989 y llegó a Júpiter en 1995. Una de las razones de que tardara tanto tiempo es que su ruta fue indirecta. Aunque la órbita de Júpiter está fuera de la de la Tierra, *Galileo* empezó dirigiéndose al interior, hacia Venus. Pasó cerca de Venus, volvió hacia la Tierra y se dirigió hacia el espacio exterior para examinar el asteroide 951 Gaspra. Luego volvió de nuevo hacia la Tierra, rodeó *otra vez* nuestro planeta hogar, y finalmente se dirigió hacia Júpiter. En el camino se acercó a otro asteroide, Ida, y descubrió que tenía su propia luna minúscula, un nuevo asteroide llamado Dactyl.

¿Por qué una trayectoria tan complicada? *Galileo* ganaba energía, y con ello velocidad, en cada uno de estos encuentros. Imaginemos una sonda espacial que se dirige hacia un planeta, no en curso de

colisión, y que se acerca mucho a la superficie del planeta, lo rodea y sale despedida al espacio. Cuando la sonda pasa por detrás del planeta ambos se atraen mutuamente. De hecho, se han estado atrayendo todo el tiempo, pero es en esta etapa cuando la fuerza de atracción alcanza su máximo y por ello tiene el mayor efecto. La gravedad del planeta da a la sonda un impulso. La energía debe conservarse, de modo que en compensación la sonda frena al planeta muy ligeramente en su órbita en torno al Sol. Puesto que la masa de la sonda es muy pequeña y la del planeta es muy grande, el efecto sobre el planeta es despreciable. El efecto sobre la sonda no lo es: puede acelerarse de forma espectacular.

Galileo llegó a menos de 16.000 kilómetros de la superficie de Venus y ganó 2,23 kilómetros por segundo en velocidad. Luego pasó a menos de 960 kilómetros de la Tierra, y otra vez a menos de 300 kilómetros, sumando otros 3,7 kilómetros por segundo. Estas maniobras eran esenciales para llegar a Júpiter, porque sus cohetes no eran lo bastante potentes para llevarle allí directamente. El plan original consistía en hacer precisamente eso, utilizando el impulsor *Centauro-G* alimentado por hidrógeno líquido. Pero el desastre de la explosión de la lanzadera *Challenger* inmediatamente después del despegue hizo que este plan se abandonara, porque el *Centauro-G* fue prohibido. Por ello *Galileo* tuvo que utilizar un impulsor más débil de combustible sólido. La misión fue un éxito enorme, y el beneficio científico incluyó la observación de la colisión entre el cometa Shoemaker-Levy 9 y Júpiter en 1994, mientras la sonda estaba aún *en ruta* hacia Júpiter.

El escenario de Xia también hace uso del efecto honda. Cuatro planetas de la misma masa forman dos pares próximos, y los miembros de cada par giran en torno a su centro de masas común en dos planos paralelos[77]. Estas raquetas de dos cuerpos juegan al tenis celeste con un quinto cuerpo, más ligero, que va y viene de una a otra en dirección perpendicular a dichos planos. El sistema está establecido de modo que cada vez que esta «pelota de tenis» pasa por un par de planetas, el efecto honda acelera la pelota y empuja el par de planetas hacia fuera a lo largo de la línea que une los dos pares, de modo que la pista de tenis se alarga y los jugadores se separan. Energía y momento se mantienen en equilibrio porque los dos planetas del par concernido se acercan ligeramente, y giran cada vez más rápidos en torno a su centro de masas. Con el montaje inicial correcto, los pares de planetas se separan cada vez con mayor rapidez, y su velocidad aumenta con tanta celeridad que llegan al infinito al cabo de un intervalo de tiempo finito. Mientras, la pelota de tenis oscila entre ellos cada vez más rápida. Los escenarios de escape de Gerver también utilizan el efecto honda.

¿Es este acto de desaparición relevante para los cuerpos celestes reales? No, si se toma literalmente. Se basa en que los cuerpos son masas puntuales. Esta es una aproximación razonable para muchos problemas en mecánica celeste, pero no lo es si los cuerpos se aproximan de modo arbitrario. Si así lo hicieran, cuerpos de tamaño finito, eventualmente colisionarían. Efectos relativistas impedirían

[77] Z. Xia, «The existence of noncollision singularities in Newtonian systems», *Annals of Mathematics*, n.º 135 (1992), pp. 411-468.

que los cuerpos se movieran a velocidad mayor que la de la luz, y cambiarían la ley de la gravedad. En cualquier caso, las condiciones iniciales, y las suposiciones de que algunas masas son idénticas, serían demasiado raras para que se den en la práctica. Sin embargo, estos ejemplos curiosos muestran que incluso si las ecuaciones de la mecánica celeste modelan muy bien la realidad en la mayoría de las circunstancias, pueden tener singularidades complicadas que impiden que existan soluciones para todo instante. También recientemente se ha comprendido que efectos honda en sistemas de estrellas triples, en los que tres estrellas orbitan unas en torno a otras en trayectorias complicadas, pueden expulsar a una de las estrellas a gran velocidad. De modo que muchas estrellas huérfanas, expulsadas de sus sistemas por sus parientes, pueden estar recorriendo la galaxia —o incluso el espacio intergaláctico— frías, solitarias, indeseadas e inadvertidas.

Cuando una ecuación diferencial se comporta de forma tan extraña que sus soluciones dejan de tener sentido al cabo de un período de tiempo finito, decimos que hay una singularidad. El trabajo anterior sobre el problema de muchos cuerpos trata realmente de varios tipos de singularidad. El problema del premio del milenio sobre la ecuación de Navier-Stokes pregunta si pueden ocurrir singularidades en problemas de valor inicial para un fluido que ocupa la totalidad del espacio o un toro plano. Si puede formarse una singularidad en un tiempo finito, es probable que el resultado sea una explosión, a menos que la propia singularidad se deshaga de alguna manera más tarde, lo que parece poco probable.

Hay dos maneras principales de acercarse a estas cuestiones. Podemos tratar de demostrar que nunca aparecen singularidades, o podemos tratar de encontrar una escogiendo condiciones iniciales adecuadas. Las soluciones numéricas pueden ayudar en los dos casos: pueden sugerir propiedades generales útiles de los flujos, y pueden proporcionar serios esperanzadores indicios sobre la posible naturaleza de las singularidades potenciales. Sin embargo, la potencial falta de exactitud en las soluciones numéricas significa que cualquiera de tales indicios debe ser tratado con cautela y justificado con más rigor.

Los intentos de demostrar la regularidad (la ausencia de singularidades) emplean una variedad de métodos para tener control sobre el flujo. Estos incluyen complicadas estimaciones de cuán grandes o pequeñas pueden hacerse ciertas variables, o técnicas más abstractas. Una aproximación popular es por vía de las denominadas soluciones débiles, que no son exactamente flujos sino estructuras matemáticas más generales con algunas de las propiedades de los flujos. Es sabido, por ejemplo, que el conjunto de singularidades de una solución débil de las ecuaciones de Navier-Stokes tridimensionales es siempre pequeño, en un sentido técnico concreto.

Se han investigado muchos escenarios diferentes que podrían llevar a singularidades. El modelo estándar de la turbulencia como una cascada de vórtices cada vez más pequeños se remonta a Andrei Kolmogorov en 1941, y él sugirió que en escalas muy pequeñas todas las formas de turbulencia tienen un aspecto muy similar. Las

proporciones de vórtices de un tamaño dado, por ejemplo, siguen una ley universal. Ahora se sabe que a medida que los vórtices se hacen más pequeños, cambian de aspecto y se hacen más largos y más finos, formando filamentos. La ley de conservación del momento angular implica que la vorticidad —con qué velocidad están girando los vórtices— debe aumentar.

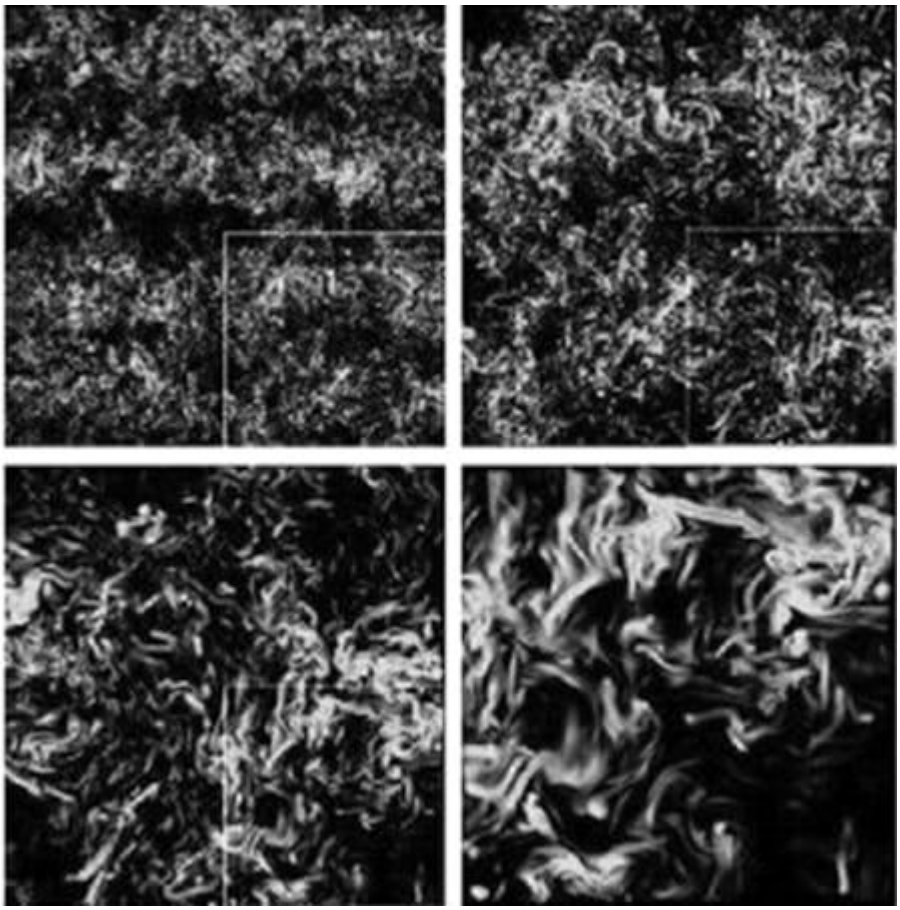


Figura 43. Zooms en un flujo turbulento, simulado con el sistema computacional VAPOR © Pablo Mininni.

Esto se llama estiramiento de vórtices, y es el tipo de comportamiento que podría dar lugar a una singularidad; por

ejemplo, si los vórtices muy pequeños pudieran hacerse infinitamente largos en tiempo finito y la vorticidad pudiera hacerse infinita en algunos puntos.

La Figura 43 muestra un *zoom* en escalas muy pequeñas de un flujo turbulento, simulado por Pablo Mininni y colegas utilizando VAPOR, la Visualization and Analysis Platform for Ocean, Atmosphere, and Solar Research (Plataforma de Visualización y Análisis para la Investigación del Océano, la Atmósfera y el Sol). Las imágenes muestran la intensidad de la vorticidad: con qué rapidez está girando el fluido. Ilustran la formación de vórtices filamentosos, las finas y delgadas estructuras en las figuras, y muestran que pueden acumularse para formar pautas a gran escala. Su programa puede realizar simulaciones en mallas cúbicas con más de tres mil millones de puntos en la malla.

En su artículo sobre este problema en la página web del Instituto Clay[78], Charles Fefferman escribe:

> *Existen muchos problemas y conjeturas fascinantes sobre el comportamiento de las soluciones de las ecuaciones de Euler y de Navier-Stokes… Puesto que ni siquiera sabemos si estas soluciones existen, nuestro conocimiento está en un nivel muy primitivo. Los métodos estándar [ecuaciones en derivadas parciales] parecen inadecuados para dirimir el problema. Es probable que, en su lugar, necesitemos algunas ideas nuevas y profundas.*

[78] http://www.claymath.org/millennium/Navier-Stokes_Equations/.

La complejidad del flujo en imágenes como las de la Figura 43 explica claramente las dificultades que es probable que encontremos cuando busquemos dichas ideas. Impertérritos, los matemáticos siguen adelante, buscando principios simples dentro de las aparentes complejidades.

Capítulo 13

Enigma cuántico

La hipótesis del hueco de masas

A pocos kilómetros al norte de Ginebra hay un recodo en la frontera entre Suiza y Francia. En la superficie, todo lo que se ve son caminos y pequeños pueblos. Pero en el subsuelo, a una profundidad de entre 50 y 175 metros, está el mayor instrumento científico del planeta. Es un gigantesco túnel circular, de más de ocho kilómetros de diámetro, unido a un segundo túnel circular de aproximadamente un cuarto del tamaño del primero. La mayor parte del mismo está bajo Francia pero dos secciones están en Suiza. Dentro de los túneles corren pares de tubos, que se cruzan en cuatro puntos.

Es el Gran Colisionador de Hadrones, cuesta 7500 millones de euros y está explorando las fronteras de la física de partículas. El objetivo principal de los diez mil científicos de los más de cien países que colaboraron en el mismo era encontrar el bosón de Higgs, o no encontrarlo, si es así como se deshizo el continuo. Lo están buscando para completar el Modelo Estándar de la física de partículas, según el cual todo lo que hay en el universo esta hecho de 17 partículas fundamentales diferentes. Según la teoría, el bosón de Higgs es lo que da masa a todas las partículas.

En diciembre de 2011 ATLAS y CMS, dos divisiones experimentales del Gran Colisionador de Hadrones, encontraron independientemente evidencia provisional de un bosón de Higgs con

una masa de unos 125 GeV (gigaelectronvoltios, unidades utilizadas en física de partículas de forma intercambiable para masa y energía, puesto que ambas son equivalentes). El 4 de julio de 2012, el CERN, el laboratorio europeo para física de partículas que controla el Gran Colisionador de Hadrones, anunció, a una abarrotada audiencia de científicos y periodistas científicos, que el continuo se había resuelto a favor del Higgs. Ambos grupos habían recogido grandes cantidades de datos adicionales, y la probabilidad de que sus datos mostraran una fluctuación aleatoria, y no una nueva partícula con propiedades tipo Higgs, había caído por debajo de 1 en 2 millones. Este es el grado de confianza que tradicionalmente se requiere en física de partículas antes de descorchar el champán.

Serán necesarios más experimentos para estar seguros de que la nueva partícula tiene todas las propiedades que debería poseer un bosón de Higgs teórico. Por ejemplo, la teoría predice que el bosón de Higgs debería tener espín 0; en el momento del anuncio, las observaciones mostraban que era o 0 o 2. Hay también una posibilidad de que «el» bosón de Higgs pueda estar compuesto de otras partículas más pequeñas, o que sea tan solo la primera de una nueva familia de partículas tipo Higgs. Así que o bien el modelo actual de las partículas fundamentales quedará reforzado, o bien tendremos nueva información que eventualmente llevará a una teoría mejor.

El último de los siete problemas del premio del milenio está íntimamente relacionado con el Modelo Estándar y el bosón de Higgs. Es una cuestión central en la teoría cuántica de campos, el

marco matemático en el que se estudia la física de partículas. Se llama la hipótesis del hueco de masas, y pone un límite inferior a la masa posible de una partícula fundamental. Es un problema representativo escogido entre una serie de grandes preguntas sin responder en esta profunda y muy nueva área de la física matemática. Tiene conexiones que van desde las fronteras de las matemáticas puras a la largo tiempo buscada unificación de las dos principales teorías físicas, la relatividad general y la teoría cuántica de campos.

En la mecánica newtoniana clásica, las magnitudes físicas básicas son espacio, tiempo y masa. El espacio se supone tridimensional euclídeo, el tiempo es una magnitud unidimensional independiente del espacio, y masa significa la presencia de materia. Las masas cambian su posición en el espacio bajo la influencia de fuerzas, y el ritmo con el que cambia su posición se mide con respecto al tiempo. La ley de movimiento de Newton describe cómo está relacionada la aceleración de un cuerpo (el ritmo de cambio de la velocidad, que a su vez es el ritmo de cambio de la posición) con la masa del cuerpo y la fuerza aplicada.

Las teorías clásicas del espacio, el tiempo y la materia alcanzaron su punto culminante en las ecuaciones de James Clerk Maxwell para el electromagnetismo[79]. Este elegante sistema de ecuaciones unificaba dos de las fuerzas de la naturaleza, que previamente se consideraban distintas. En lugar de electricidad y magnetismo, había un único campo electromagnético. Un campo llena la

[79] Véase Ian Stewart, *Seventeen Equations that Changed the World*, Profile, 2012, capítulo 14.

totalidad del espacio, como si el universo estuviera lleno de cierto tipo de fluido invisible. En cada punto del espacio podemos medir la intensidad y dirección del campo, como si dicho fluido estuviera fluyendo con pautas matemáticas. Para algunos fines el campo electromagnético puede separarse en dos componentes, el campo eléctrico y el campo magnético. Pero un campo magnético en movimiento crea un campo eléctrico, y recíprocamente, de modo que cuando se llega a la dinámica, ambos campos deben combinarse en un único campo más complejo.

Esta confortable imagen del mundo físico, en la que los conceptos científicos fundamentales guardan un estrecho parecido con cosas que perciben nuestros sentidos, cambió drásticamente en los primeros años del siglo XX. En ese momento los físicos empezaron a darse cuenta de que en escalas muy pequeñas, demasiado pequeñas para ser observadas con cualquier microscopio entonces disponible, la materia es muy diferente de lo que todos habían imaginado. Físicos y químicos empezaron a tomar en serio una teoría muy especulativa que se retrotraía más de dos milenios hasta las elucubraciones filosóficas de Demócrito en la antigua Grecia y otros estudiosos en la India. Era la idea de que aunque el mundo parece estar hecho de incontables materiales diferentes, toda la materia está formada de partículas minúsculas: los átomos. La palabra procede del término griego que designa «indivisible».

Los químicos del siglo XIX encontraron evidencia indirecta a favor de los átomos: los elementos que se combinan para formar moléculas más complejas lo hacen en proporciones muy específicas,

a menudo próximas a números enteros. John Dalton formuló estas observaciones en su ley de las proporciones múltiples, y propuso los átomos como una explicación. Si cada compuesto químico consistía de números fijos de átomos de varios tipos, una proporción de este tipo aparecería de forma automática. Por ejemplo, ahora sabemos que cada molécula de dióxido de carbono consiste en dos átomos de oxígeno y un átomo de carbono, de modo que los números de átomos estarán en razón de dos a uno. Sin embargo, hay complicaciones: átomos diferentes tienen masas diferentes, y muchos elementos se presentan como moléculas formadas por varios átomos. Por ejemplo, la molécula de oxígeno está compuesta de dos átomos de oxígeno. Si no nos damos cuenta de lo que está pasando, pensaríamos que un átomo de oxígeno es de una masa doble de la que tiene en realidad. Y algunos elementos aparentes son en realidad mezclas de diferentes «isótopos» (estructuras atómicas). Por ejemplo, el cloro se da en la naturaleza como una mezcla de dos formas estables, ahora llamadas cloro 35 y cloro 37, en proporciones de alrededor de un 76 por 100 y un 24 por 100, respectivamente. Por eso, el «peso atómico» observado es 35,45, que en las etapas iniciales de la teoría atómica se interpretaba en el sentido de que «el átomo de cloro está compuesto de treinta y cinco y medio átomos de hidrógeno». Y eso significa que un átomo no es indivisible. Cuando se iniciaba el siglo XX la mayoría de los científicos seguían pensando que el salto a la teoría atómica era demasiado grande, y la evidencia numérica era demasiado débil para justificarlo.

Algunos científicos, en especial Maxwell y Ludwig Boltzmann, fueron más allá, convencidos de que los gases son colecciones de moléculas tenuemente distribuidas y que las moléculas están hechas por ensamblaje de átomos. Lo que, al parecer, convenció a sus colegas fue la explicación que dio Albert Einstein para el movimiento browniano, los erráticos movimientos de minúsculas partículas suspendidas en un fluido que eran visibles al microscopio. Einstein decidió que estos movimientos debían estar causados por colisiones con moléculas del fluido que se movían de modo aleatorio y realizó algunos cálculos cuantitativos para apoyar esa idea. Jean Perrin confirmó experimentalmente estas predicciones en 1908. Ser capaces de ver el efecto de las supuestas partículas indivisibles de materia, y de hacer predicciones cuantitativas, aportaba más convicción que las elucubraciones filosóficas y la numerología curiosa. En 1911 ya había un consenso científico en la existencia de los átomos.

Mientras pasaba esto, unos pocos científicos empezaron a darse cuenta de que los átomos no son indivisibles. Tienen algún tipo de estructura, y es posible separar pequeños fragmentos de ellos. En 1897 Joseph John Thomson estaba experimentando con los denominados rayos catódicos y descubrió que podía hacerse que los átomos emitieran partículas todavía más minúsculas, los electrones. No solo eso: átomos de elementos diferentes emitían las mismas partículas. Aplicando un campo magnético, Thomson demostró que los electrones llevan una carga eléctrica negativa. Puesto que un átomo es eléctricamente neutro, debe haber también una parte de

los átomos con una carga positiva, lo que llevó a Thomson a proponer el modelo del pudin de pasas: un átomo es como un pudín cargado positivamente salpicado con pasas cargadas negativamente. Pero en 1909 uno de los ex-estudiantes de Thomson, Ernest Rutherford, realizó experimentos que mostraban que la mayor parte de la masa de un átomo está concentrada cerca de su centro. Los pudines no son así.

¿Cómo pueden los experimentos explorar regiones tan minúsculas del espacio? Imaginemos una parcela de tierra, que puede tener o no edificios u otras estructuras. No se nos permite entrar en el área, que además está oscura de modo que no podemos ver lo que hay allí. Sin embargo, tenemos un rifle y muchas cajas de munición. Podemos disparar balas aleatoriamente a la parcela y observar en qué dirección salen. Si la parcela es como un pudin de pasas, la mayoría de las balas la atravesarán directamente. Si en ocasiones tenemos que agacharnos cuando una bala rebota hacia nosotros, es que hay algo muy sólido en alguna parte. Observando con qué frecuencia la bala sale a un ángulo dado, podemos estimar el tamaño del objeto sólido.

Las balas de Rutherford eran partículas alfa, núcleos de átomos de helio, y su parcela de tierra era una fina hoja de oro. El trabajo de Thomson había mostrado que las pasas-electrones tenían una masa muy pequeña, de modo que casi toda la masa de un átomo debería encontrarse en el pudin. Si el pudin no tenía grumos, la mayoría de las partículas alfa deberían atravesarlo directamente, y muy pocas serían desviadas y no mucho. En lugar de ello, una proporción

pequeña pero significativa experimentaba grandes desviaciones. Así que la imagen del pudin de pasas no funcionaba. Rutherford sugirió una metáfora diferente, una metáfora que todavía utilizamos hoy de manera informal pese a que ha sido superada por imágenes más modernas: el modelo planetario. Un átomo es como un sistema solar; tiene un enorme núcleo central, su sol, alrededor del cual orbitan los electrones como planetas. Así pues, como el Sistema Solar, el interior de un átomo es básicamente espacio vacío.

Rutherford llegó a encontrar pruebas de que el núcleo está compuesto de dos tipos diferentes de partículas: protones, con carga positiva; y neutrones, con carga cero. Los dos tienen masas muy parecidas, y ambos son unas mil ochocientas veces más masivos que un electrón. Los átomos, lejos de ser indivisibles, están hechos de partículas subatómicas aún más pequeñas. Esta teoría explica la numerología en enteros de los elementos químicos: lo que se está contando son los números de protones y neutrones. También explica los isótopos: sumar o restar algunos neutrones cambia la masa, pero mantiene la carga total cero y deja invariable el número de electrones, que es igual al número de protones. Las propiedades químicas de un átomo están básicamente controladas por sus electrones. Por ejemplo, el cloro 35 tiene 17 protones, 17 electrones y 18 neutrones; el cloro 37 tiene 17 protones, 17 electrones y 20 neutrones. La cifra 35,45 aparece porque el cloro natural es una mezcla de estos dos isótopos.

A comienzos del siglo XX había una nueva teoría en juego, aplicable a la materia en las escalas de las partículas subatómicas. Era la

mecánica cuántica, y una vez que estuvo disponible, la física ya no volvería a ser la misma. La mecánica cuántica predecía una gran cantidad de fenómenos nuevos, muchos de los cuales fueron en seguida observados en el laboratorio. Explicaba muchas observaciones extrañas y que previamente resultaban desconcertantes. Predecía la existencia de nuevas partículas fundamentales. Y nos decía que la imagen clásica del universo en que vivimos, y que hasta entonces había tenido un excelente acuerdo con las observaciones, es falsa. Nuestras percepciones a escala humana son pobres modelos de la realidad en su nivel más fundamental.

En física clásica la materia está hecha de partículas y la luz es una onda. En mecánica cuántica la luz es también una partícula: el fotón. De modo recíproco, la materia, por ejemplo los electrones, puede comportarse a veces como una onda. La previamente nítida divisoria entre ondas y partículas no solo se difuminaba sino que desaparecía por completo, reemplazada por la dualidad onda/partícula. El modelo planetario del átomo no funcionaba muy bien si se tomaba al pie de la letra, de modo que apareció una nueva imagen. En lugar de orbitar en torno al núcleo como planetas, los electrones forman una nube difusa centrada en el núcleo, una nube no de materia sino de probabilidad. La densidad de la nube corresponde a la probabilidad de encontrar un electrón en dicha localización.

Además de protones, neutrones y electrones, los físicos conocían otra partícula subatómica, el fotón. Pronto aparecieren otras. Un

fallo aparente de la ley de conservación de la energía llevó a Wolfgang Pauli a proponer un arreglo provisional que postulaba la existencia del neutrino, una nueva partícula invisible y prácticamente indetectable que proporcionaría la energía que faltaba. Era lo bastante detectable para que se confirmara su existencia en 1956. Y eso abrió las compuertas. Pronto había piones, muones y kaones, los últimos descubiertos al observar rayos cósmicos. Había nacido la física de partículas, que siguió utilizando el método de Rutherford para explorar las increíblemente minúsculas escalas espaciales implicadas: para descubrir lo que hay en el interior de algo, se arroja un montón de materia contra ello y se observa lo que rebota. Se construyeron y pusieron en operación aceleradores de partículas cada vez más grandes, las pistolas que disparaban las balas. El acelerador lineal de Stanford tenía tres kilómetros de longitud. Para no tener que construir aceleradores cuyas longitudes abarcaran continentes, se curvaron en círculo, de modo que las partículas pudiesen dar muchísimas vueltas a velocidades enormes. Eso complicaba la tecnología, porque las partículas que se mueven en círculos radian energía, pero había remedios.

El primer fruto de estos trabajos fue un cada vez mayor catálogo de partículas supuestamente fundamentales. Enrico Fermi expresaba así su frustración: «Si pudiera recordar los nombres de todas estas partículas, sería un botánico». Pero de cuando en cuando nuevas ideas procedentes de la teoría cuántica reducían la lista, y se

proponían nuevos tipos de partículas cada vez más pequeñas para unificar las estructuras ya observadas.

La mecánica cuántica primitiva se aplicaba a objetos individuales tipo onda o tipo partícula. Pero inicialmente nadie podía describir un buen análogo mecano-cuántico de un campo. Era imposible ignorar esta laguna porque las partículas (descriptibles por la mecánica cuántica) podían interaccionar y lo hacían con campos (no descriptibles por la mecánica cuántica). Era como querer descubrir cómo se mueven los planetas del Sistema Solar si se conocieran las leyes de movimiento de Newton (cómo se mueven las masas cuando se aplican fuerzas), pero no se conociera su ley de la gravedad (cuáles son dichas fuerzas).

Había otra razón para querer modelar los campos y no solo las partículas. Gracias a la dualidad onda/partícula, están íntimamente relacionadas. Una partícula es, en esencia, un grumo en un campo. Un campo es un mar repleto de partículas. Los dos conceptos son inseparables. Por desgracia, los métodos desarrollados hasta esa fecha se basaban en que las partículas son como puntos minúsculos, y no se extendían a los campos de ninguna manera razonable. No se podían pegar montones de partículas y llamar campo al resultado, porque las partículas *interaccionan* unas con otras.

Imaginemos una multitud en… bueno, un campo. Quizá están en un concierto de *rock*. Vista desde un helicóptero, la multitud se parece a un fluido, chapoteando por el campo —a menudo literalmente, por ejemplo en el Festival Glastonbury, reputado por

convertirse en un mar de lodo—. Desde el suelo se hace claro que el fluido es en realidad una masa agitada de partículas individuales: personas. O quizá densos racimos de personas, como algunos amigos que pasean juntos, formando una unidad indivisible, o como un grupo de extraños que llega con un propósito común, tal como ir al bar. Pero no se puede modelar adecuadamente la multitud sumando lo que harían las personas si cada una fuera a lo suyo. Cuando un grupo se dirige al bar, bloquea el camino de otro grupo. Los dos grupos colisionan y se agolpan. Establecer una teoría cuántica de campos efectiva es como hacer esto cuando las personas son funciones de onda cuánticas localizadas.

A finales de los años veinte del siglo pasado, razonamientos de este tipo habían convencido a los físicos de que, por difícil que pudiera ser la tarea, la mecánica cuántica tenía que ser ampliada para tener en cuenta a los campos tanto como a las partículas. El lugar natural para empezar era el campo electromagnético. De algún modo los componentes eléctrico y magnético de este campo tenían que ser cuantificados: reescritos en el formalismo de la mecánica cuántica. Matemáticamente, este formalismo era poco familiar y no muy físico. Los observables —cosas que se podían medir— ya no se representaban utilizando los viejos números. En su lugar, correspondían a operadores en un espacio de Hilbert: reglas matemáticas para manipular ondas. Estos operadores violaban las hipótesis habituales de la mecánica clásica. Si se multiplican dos números, el resultado es el mismo sea cual sea el primero. Por ejemplo, 2×3 y 3×2 dan lo mismo. Esta propiedad, llamada

conmutatividad, falla para muchos pares de operadores, igual que ponerse primero los calcetines y luego los zapatos no tiene el mismo efecto que ponerse primero los zapatos y luego los calcetines. Los números son criaturas pasivas, los operadores son activos. Qué acción se realiza primero fija el escenario para la otra.

La conmutatividad es una propiedad matemática muy agradable. Su ausencia es algo molesta, y esta es precisamente una de las razones por las que cuantizar un campo resulta ser complicado. No obstante, a veces puede hacerse. El campo electromagnético fue cuantizado en una serie de etapas, empezando con la teoría del electrón de Dirac en 1928, y completadas por Sin-Itiro Tomonaga, Julian Schwinger, Richard Feynman y Freeman Dyson a finales de los años cuarenta y principios de los cincuenta. La teoría resultante se conoce como electrodinámica cuántica.

El punto de vista que se utilizaba allí se basaba en un método que podría trabajar de forma más general. La idea subyacente se remontaba directamente hasta Newton. Cuando los matemáticos intentaron resolver las ecuaciones que proporcionaba la ley de Newton descubrieron algunos trucos generales y útiles, conocidos como leyes de conservación. Cuando se mueve un sistema de masas, algunas magnitudes permanecen constantes. La más familiar es la energía, que viene en dos sabores, cinética y potencial. La energía cinética está relacionada con la velocidad con que se mueve un cuerpo, y la energía potencial es el trabajo hecho por las fuerzas. Cuando se deja caer una piedra desde el borde de un acantilado intercambia energía potencial, debida a la gravedad, por

energía cinética; dicho en lenguaje ordinario, cae y se acelera. Otras magnitudes conservadas son el momento lineal, que es el producto de la masa por la velocidad, y el momento angular, que está relacionado con el ritmo de giro de un cuerpo. Estas magnitudes conservadas relacionan las diferentes variables utilizadas para describir el sistema, y por consiguiente reducen su número. Eso ayuda cuando se trata de resolver las ecuaciones, como vimos en el caso del problema de dos cuerpos en el capítulo 8.

En el siglo XX se había descubierto la fuente de estas leyes de conservación. Emmy Noether demostró que toda magnitud conservada corresponde a un grupo continuo de simetrías de las ecuaciones. Una simetría es una transformación matemática que deja las ecuaciones invariables, y todas las simetrías forman un grupo, para el que la operación de composición es «hacer una transformación y luego la otra». Un grupo continuo es un grupo de simetrías definidas por un único número real. Por ejemplo, la rotación en torno a un eje dado es una simetría, y el ángulo de rotación puede ser cualquier número real, de modo que las rotaciones —de cualquier ángulo— en torno a un eje dado forman una familia continua. Aquí la magnitud conservada asociada es el momento angular. Análogamente, el momento lineal es la magnitud conservada asociada a la familia de traslaciones en una dirección dada. ¿Qué pasa con la energía? Esta es la magnitud conservada que corresponde a las simetrías temporales; las ecuaciones son las mismas en cualquier instante de tiempo.

Cuando los físicos trataron de unificar las fuerzas básicas de la naturaleza, llegaron a convencerse de que las simetrías eran la clave. La primera de estas unificaciones fue la de Maxwell, que combinaba electricidad y magnetismo en un único campo electromagnético. Maxwell consiguió esta unificación sin considerar la simetría, pero pronto se hizo claro que sus ecuaciones poseen un notable tipo de simetría que no había sido advertida previamente: la simetría gauge. Y eso parecía una palanca estratégica que podría abrir teorías cuánticas de campos más generales.

Rotaciones y traslaciones son simetrías globales: se aplican uniformemente a lo largo de todo el espacio y el tiempo. Una rotación en torno a un eje rota todo punto del espacio el mismo ángulo. Las simetrías gauge son diferentes: son simetrías locales, que pueden variar de un punto a otro en el espacio. En el caso del electromagnetismo, estas simetrías locales son cambios de fase. Una oscilación local del campo electromagnético tiene una amplitud (qué tamaño tiene) y una fase (el momento en que alcanza su máximo). Si tomamos una solución de las ecuaciones de campo de Maxwell y cambiamos la fase en cada punto, obtenemos otra solución, siempre que hagamos un cambio compensatorio en la descripción del campo incorporando una carga electromagnética local.

Las simetrías gauge fueron introducidas por Hermann Weyl en un intento infructuoso por conseguir otra unificación, la del electromagnetismo y la relatividad general; es decir, las fuerzas electromagnéticas y gravitatorias. El nombre vino de un equívoco: él pensaba que las simetrías locales correctas deberían ser cambios de

escala espacial, o «gauge[iii]». Esta idea no funcionó, pero el formalismo de la mecánica cuántica llevó a Vladimir Fock y Fritz London a introducir un tipo diferente de simetría local. La mecánica cuántica se formula utilizando números complejos, no solo números reales, y toda función de onda cuántica tiene una fase compleja. Las simetrías locales relevantes rotan la fase cualquier ángulo en el plano complejo. En abstracto, este grupo de simetrías consiste en todas las rotaciones, pero en coordenadas complejas estas son «transformaciones unitarias» (U) en un espacio con una dimensión compleja (1), de modo que el grupo formado por estas simetrías se denota por U(1). Aquí el formalismo no es solo un juego matemático abstracto: permite a los físicos escribir, y luego resolver, las ecuaciones para partículas cuánticas cargadas que se mueven en un campo electromagnético. En las manos de Tomonaga, Schwinger, Feynman y Dyson este punto de vista llevó a la primera teoría cuántica de campos relativista del electromagnetismo: la electrodinámica cuántica. La simetría bajo el grupo gauge U(1) fue fundamental para su trabajo.

El siguiente paso, que unifica la electrodinámica cuántica con la fuerza nuclear débil, fue conseguido por Abdus Salam, Sheldon Glashow, Steven Weinberg y otros en los años sesenta del siglo pasado. Junto al campo electromagnético con su simetría gauge U(1), introdujeron campos asociados con cuatro partículas fundamentales, los denominados bosones W^+, W^0, W^- y B^0. Las simetrías gauge de este campo, que en efecto rotan combinaciones de estas partículas para producir otras combinaciones, forman otro

grupo, llamado SU(2) —transformaciones unitarias (U) en un espacio complejo bidimensional (2) que son también especiales (S), una condición técnica sencilla—. Por lo tanto el grupo gauge combinado es U(1)×SU(2), donde el×indica que los dos grupos actúan independientemente sobre los dos campos. El resultado, llamado teoría electrodébil, requería una difícil innovación matemática. El grupo U(1) para la electrodinámica cuántica es conmutativo: aplicar sucesivamente dos transformaciones de simetría da el mismo resultado, cualquiera que sea el orden en que se aplican. Esta agradable propiedad simplifica mucho las matemáticas, pero no es válida para SU(2). Esta fue la primera aplicación de una teoría gauge no conmutativa.

La fuerza nuclear fuerte entra en juego cuando consideramos la estructura interna de partículas como protones y neutrones. El gran avance en esta área fue motivado por una curiosa pauta matemática en una clase concreta de partículas, llamadas hadrones. La pauta era conocida como el óctuple camino. Inspiró la teoría de la cromodinámica cuántica, que postulaba la existencia de partículas ocultas llamadas *quarks* y las utilizaba como componentes básicos para el gran zoo de hadrones.

En el modelo estándar, todo lo que hay en el universo depende de dieciséis partículas genuinamente fundamentales, cuya existencia ha sido confirmada por experimentos en aceleradores. Más una decimoséptima, que actualmente está buscando el Gran Colisionador de Hadrones. De las partículas conocidas para Rutherford, solo dos siguen siendo fundamentales: el electrón y el

fotón. El protón y el neutrón, por el contrario, están hechos de *quark*s. El nombre fue acuñado por Murray Gell-Mann, que pretendía que rimara con «cork». Él dio con un pasaje del *Finnegans Wake* de James Joyce:

Three quarks for Muster Mark!
Sure he has not got much of a bark
And sure any he has it's all beside the mark.

Esto parecería apuntar a una pronunciación que rima con «mark», pero Gell-Mann encontró una manera de justificar su intención. Ambas pronunciaciones son ahora habituales.

El Modelo Estándar considera seis *quark*s, dispuestos en pares. Tienen nombres curiosos: *up*/*down, charmed*/*strange*, y *top*/*bottom*. Hay seis leptones, también en pares: el electrón, el muón y el tauón (hoy llamado simplemente tau) y sus neutrinos asociados. Estas doce partículas son colectivamente llamadas fermiones, en referencia a Fermi. Las partículas se mantienen unidas gracias a fuerzas, que son de cuatro tipos: gravedad, electromagnetismo, la fuerza nuclear fuerte y la fuerza nuclear débil. Dejando aparte la gravedad, que todavía no ha sido plenamente reconciliada con la imagen cuántica, esto da tres fuerzas. En física de partículas, las fuerzas se producen por un intercambio de partículas, que son «portadoras» o «mediadoras» de la fuerzas. La analogía habitual es con dos jugadores de tenis que se mantienen unidos por su mutua atención a la bola. El fotón es el portador de la fuerza electromagnética, los bosones Z y W son portadores de la fuerza

nuclear débil, y los gluones son portadores de la fuerza nuclear fuerte. Mejor dicho, los gluones son técnicamente los portadores de la fuerza de color que mantiene unidos a los *quarks*, y la fuerza fuerte es lo que observamos como resultado.

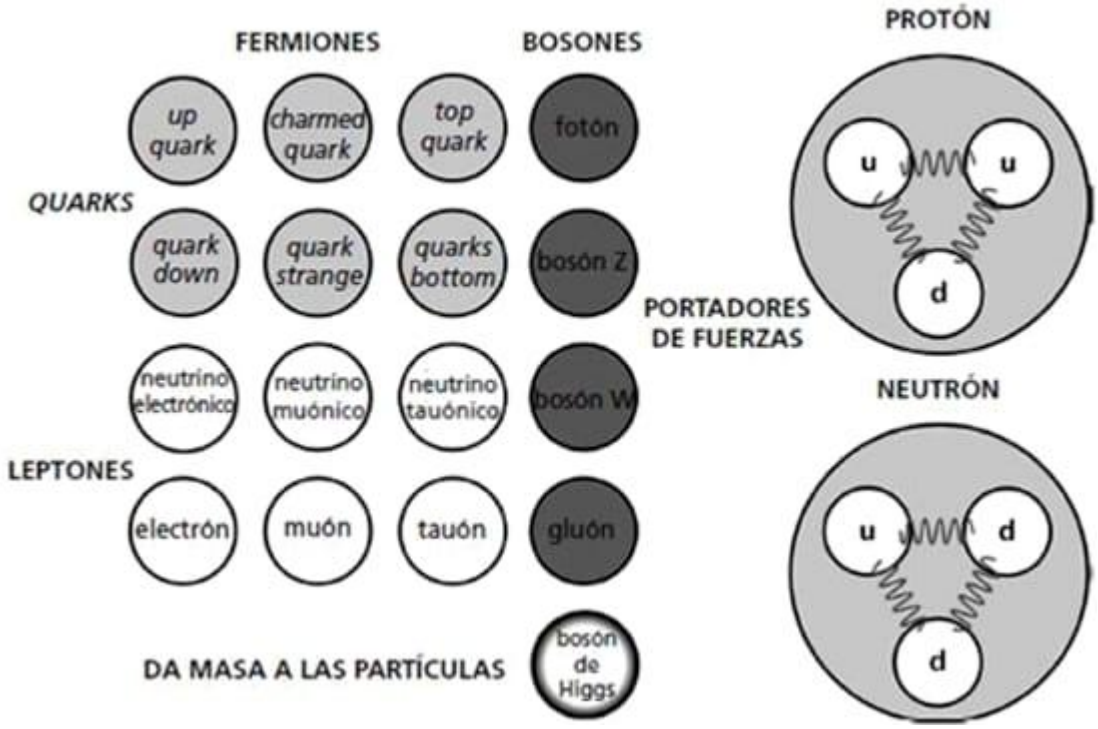


Figura 44. Izquierda: Las 17 partículas del Modelo Estándar. Derecha: Cómo hacer un protón y un neutrón a partir de quarks. Derecha, arriba: Protón = dos quarks up + un quark down. Derecha, abajo: Neutrón = un quark up + dos quarks down.

El protón consiste en dos *quarks up* más un *quark down*; el neutrón consta de un *quark up* más dos *quarks down*. En cada una de estas partículas los *quarks* se mantienen unidos por los gluones. Estos cuatro portadores de fuerzas se conocen colectivamente como bosones, en referencia a Chandra Bose. La distinción entre

fermiones y bosones es importante: tienen diferentes propiedades estadísticas. La Figura 44 (*izquierda*) muestra el catálogo resultante de partículas que se conjeturan fundamentales. La Figura 44 (*derecha*) muestra cómo hacer un protón y un neutrón a partir de *quarks*.

El bosón de Higgs completa esta imagen al explicar por qué las otras 16 partículas del Modelo Estándar tienen masas no nulas. Recibe su nombre de Peter Higgs, uno de los físicos que sugirieron la idea. Otros físicos involucrados son Philip Anderson, François Englert, Robert Brout, Gerald Guralnik, Carl Hagen y Thomas Kibble. El bosón de Higgs es la encarnación en forma de partícula de un hipotético campo cuántico, el campo de Higgs, con una propiedad inusual pero vital: en un vacío, el campo es distinto de cero. Las otras 16 partículas están influidas por el campo de Higgs, que hace que se comporten como si tuvieran masa.

En 1993 David Miller, en respuesta a un reto del ministro de Ciencia británico William Waldegrave, presentó una analogía sorprendente: una fiesta. Los asistentes están repartidos uniformemente por la habitación cuando entra el invitado de honor (una ex primera ministra). Al instante, todos se agrupan a su alrededor. A medida que ella se mueve por el salón, unas personas se unen al grupo y otras lo dejan, y el grupo en movimiento le da a ella una masa extra, que hace más difícil que ella se detenga. Este es el mecanismo de Higgs. Imaginemos ahora que circula un rumor por el salón y la gente se agolpa para oír la noticia. Este grupo es el bosón de Higgs. Miller añadía: «Podría haber un mecanismo de

Higgs, y un campo de Higgs que llenase nuestro universo, sin que haya un bosón de Higgs. La próxima generación de colisionadores lo resolverá». Parece que ahora se ha resuelto el bosón de Higgs, pero el campo de Higgs aún necesita más trabajo.

La cromodinámica cuántica es otra teoría gauge, esta vez con grupo gauge SU(3). Como sugiere la notación, la transformación actúa ahora sobre el espacio complejo tridimensional. Luego siguió la unificación del electromagnetismo, la fuerza débil y la fuerza fuerte. Supone la existencia de tres campos cuánticos, uno por cada fuerza, con grupos gauge U(1), SU(2) y SU(3) respectivamente. La combinación de los tres da el Modelo Estándar, con grupo gauge U(1)×SU(2)×SU(3). Siendo rigurosos, las simetrías SU(2) y SU(3) son aproximadas; se cree que se hacen exactas a energías muy altas. Por eso su efecto sobre las partículas que forman nuestro mundo corresponde a simetrías «rotas», las trazas de estructura que permanecen cuando el sistema ideal perfectamente simétrico está sujeto a una pequeña perturbación.

Los tres grupos contienen familias continuas de simetrías: una familia para U(1), tres para SU(2) y ocho para SU(3). Asociadas a ellas hay varias magnitudes conservadas. Las simetrías de la mecánica newtoniana proporcionan de nuevo energía, momento lineal y momento angular. Las magnitudes conservadas para las simetrías gauge U(1)×SU(2)×SU(3) son varios «números cuánticos» que caracterizan las partículas. Son análogas a magnitudes tales como espín y carga, pero se aplican a los *quarks*; tienen nombres como carga de color, isospín e hipercarga. Finalmente, hay algunas

magnitudes conservadas adicionales para U(1): números cuánticos para los seis leptones, tales como número electrónico, número muónico y número tauónico. El resultado es que las simetrías de las ecuaciones del Modelo Estándar, vía teorema de Noether, explican las variables físicas nucleares de las partículas fundamentales.

El mensaje importante para nuestra historia es la estrategia y el resultado general. Para unificar las teorías físicas hay que encontrar sus simetrías y unificarlas. Luego hay que idear una teoría adecuada con dicho grupo de simetrías combinado. No estoy sugiriendo que el proceso sea directo; de hecho es técnicamente muy complejo. Pero así es como se ha desarrollado la teoría cuántica de campos hasta ahora, y solo una de las cuatro fuerzas de la naturaleza cae hoy día fuera de su alcance: la gravedad.

El teorema de Noether no solo explica las principales variables físicas asociadas con partículas fundamentales: fue así como se encontraron muchas de las simetrías subyacentes. Los físicos trabajaron hacia atrás, a partir de los números cuánticos observados e inferidos, para deducir qué simetrías debería tener el modelo. Luego escribieron ecuaciones adecuadas con dichas simetrías y confirmaron que dichas ecuaciones encajaban muy estrechamente la realidad. De momento, este paso final requiere escoger los valores de 19 parámetros, números que deben ser introducidos en las ecuaciones para dar resultados cuantitativos. Nueve de estos son masas de partículas concretas: los seis *quarks* y el electrón, el muón y el tau. Los demás son más técnicos, cosas como ángulos de mezcla y acoplamientos de fase. Diecisiete de estos

parámetros son conocidos gracias a los experimentos, pero dos no lo son; describen el todavía hipotético *campo* de Higgs. Pero ahora hay buenas perspectivas de medirlos, porque los físicos saben dónde mirar.

Las ecuaciones utilizadas en estas teorías pertenecen a una clase general de teorías de campos gauge, conocidas como teorías de Yang-Mills. En 1954 Chen-Ning Yang y Robert Mills intentaron desarrollar teorías gauge para explicar la fuerza fuerte y las partículas asociadas a ella. Sus primeros intentos tropezaban con dificultades cuando se cuantizaba el campo, porque esto exigía que las partículas tuvieran masa cero. En 1960 Jeffrey Goldstone, Yoichiro Nambu y Giovanni Jona-Lasinio encontraron una manera de evitar este problema: empezar con una teoría que predecía partículas sin masa, pero luego modificarla rompiendo alguna de las simetrías. Es decir, cambiar un poco las ecuaciones introduciendo nuevos términos asimétricos. Cuando se utilizó esta idea para modificar la teoría de Yang-Mills, las ecuaciones resultantes funcionaban muy bien, tanto en la teoría electrodébil como en cromodinámica cuántica.

Yang y Mills suponían que el grupo gauge era un grupo unitario especial. Para las aplicaciones a partículas este era el SU(2) o el SU(3), el grupo unitario especial para dos o tres dimensiones complejas, pero el formalismo funcionaba en cualquier número de dimensiones. Su teoría aborda en línea directa una difícil pero inevitable dificultad matemática. El campo electromagnético es engañosamente simple en un aspecto: sus simetrías gauge

conmutan. A diferencia de la mayoría de los operadores cuánticos, el orden en que se cambian las fases no afecta a las ecuaciones. Pero ahora los físicos tenían puesto los ojos en una teoría cuántica de campos para partículas subatómicas. Aquí, el grupo gauge era no conmutativo, lo que hacía muy difícil cuantizar las ecuaciones.

Yang y Mills lo consiguieron utilizando una representación diagramática de interacciones de partículas introducida por Richard Feynman. Cualquier estado cuántico puede considerarse como una superposición de innumerables interacciones de partículas. Por ejemplo, incluso un vacío incluye pares de partículas y antipartículas que en cada instante nacen e inmediatamente desaparecen. Una simple colisión entre dos partículas se divide en una danza desconcertante de apariciones y desapariciones de partículas intermediarias, yendo y viniendo, dividiéndose y combinándose. Lo que salva la situación es una combinación de dos cosas. Las ecuaciones de campo pueden cuantizarse para cada diagrama de Feynman concreto y todas estas contribuciones pueden sumarse para representar el efecto global de la interacción. Además, los diagramas más complicados rara vez aparecen, de modo que su contribución a la suma no es muy grande. Incluso así, hay un serio problema. La suma, interpretada directamente, es infinita. Yang y Mills encontraron una manera de «renormalizar» el cálculo de modo que se eliminaba una infinidad de términos que no deberían contar de hecho. Lo que quedaba era una suma finita, y su valor encajaba muy bien con la realidad. Esta técnica era completamente misteriosa cuando se ideó por primera vez, pero ahora tiene sentido.

En los años setenta entraron en escena los matemáticos, y Michael Atiyah generalizó la teoría de Yang-Mills a una amplia clase de grupos gauge. Matemáticas y física empezaron a realimentarse mutuamente, y el trabajo de Edward Witten y Nathan Seiberg sobre teorías cuánticas de campos topológicas llevó al concepto de supersimetría, en donde todas las partículas conocidas tienen nuevas contrapartidas «supersimétricas»: electrones y selectrones, *quarks* y *squarks*. Esto simplificaba las matemáticas y llevaba a predicciones físicas. Sin embargo, estas nuevas partículas todavía no han sido observadas, y es probable que algunas deberían manifestarse ahora en los experimentos realizados utilizando el Gran Colisionador de Hadrones. El valor matemático de estas ideas está bien establecido, pero su relevancia directa para la física no lo está. Sin embargo, arrojan luz útil sobre la teoría de Yang-Mills.

La teoría cuántica de campos es una de las fronteras en más rápido movimiento de la física matemática, y por ello el Instituto Clay quería incluir algo relativo a este tema como uno de los premios del milenio. La hipótesis del hueco de masas se sitúa en el centro de esta rica área y aborda una importante cuestión matemática vinculada con la física de partículas. La aplicación de los campos de Yang-Mills para describir las partículas fundamentales en términos de la fuerza nuclear fuerte depende de una característica específica de la teoría cuántica conocida como un hueco de masas. En relatividad, una partícula que viaja a la velocidad de la luz adquiere una masa infinita, a menos que su masa en reposo sea nula. El hueco de masas permite que las partículas cuánticas tengan masas

no nulas finitas, incluso si las ondas clásicas asociadas viajan a la velocidad de la luz. Cuando existe un hueco de masas, cualquier estado que no sea el vacío tiene una energía que supera a la del vacío en al menos una cantidad fija. Es decir, hay un límite inferior no nulo para la masa de una partícula.

Los experimentos confirman la existencia de un hueco de masas, y las simulaciones por ordenador de las ecuaciones apoyan la hipótesis del hueco de masas. Sin embargo, no podemos suponer que un modelo encaja con la realidad y luego utilizar la realidad para verificar propiedades matemáticas del modelo, porque caeríamos en un círculo vicioso. Por eso es necesaria una comprensión teórica. Un paso clave sería una demostración rigurosa de que existen versiones cuánticas de la teoría de Yang-Mills. La versión clásica (no cuántica) está hoy bastante bien entendida, pero la análoga cuántica está viciada, afectada por el problema de la renormalización, esos molestos infinitos que tienen que ser neutralizados mediante malabarismos matemáticos.

Un enfoque atractivo empieza por convertir el espacio continuo en una red discreta y escribir un análogo en la red a la ecuación de Yang-Mills. Entonces la cuestión principal es mostrar que a medida que la red se hace cada vez más fina, y se aproxima a un continuo, este análogo converge a un objeto matemático bien definido. Algunas propiedades necesarias de las matemáticas pueden inferirse por intuición física, y si estas propiedades pudieran establecerse rigurosamente sería posible demostrar que existe una adecuada teoría de Yang-Mills cuántica. La hipótesis del hueco de

masas implica una comprensión más detallada de cómo las teorías reticulares se aproximan a esta hipotética teoría de Yang-Mills. Por lo tanto, la existencia de la teoría, y la hipótesis del hueco de masas, están estrechamente entretejidas.

En eso es en donde todos están atascados. En 2004 Michael Douglas escribió un informe sobre el estatus del problema, y dijo: «Hasta donde yo sé, no se ha hecho ningún avance importante sobre este problema en los últimos años. En particular, aunque se han hecho progresos en teorías de campos en dimensiones menores, no conozco ningún progreso importante hacia una construcción matemáticamente rigurosa de la teoría de Yang-Mills cuántica». Esta evaluación sigue pareciendo correcta.

Sin embargo, se han hecho progresos más impresionantes en algunos problemas relacionados que pueden arrojar luz útil. Teorías cuánticas de campos especiales, conocidas como modelos sigma bidimensionales, son más tratables, y la hipótesis del hueco de masas ha sido establecida para uno de estos modelos. Las teorías cuánticas de campos supersimétricas, que incluyen hipotéticas supercompañeras de las partículas fundamentales habituales, tienen propiedades matemáticas agradables que, en efecto, eliminan la necesidad de renormalización. Físicos como Edward Witten han estado haciendo progresos hacia la solución de cuestiones relacionadas en el caso supersimétrico. Hay esperanzas de que algunos de los métodos que surgen de este trabajo pudieran sugerir nuevas maneras de abordar el problema original. Pero cualesquiera que puedan ser las implicaciones físicas, y cualquiera que sea el

estatus final de la hipótesis del hueco de masas, muchos de estos desarrollos ya han enriquecido las matemáticas con nuevos conceptos y nuevas herramientas importantes.

Capítulo 14

Sueños diofánticos

La conjetura de Birch-Swinnerton-Dyer

En el capítulo 7 encontramos la *Arithmetica* de Diofanto, y comenté que seis de sus trece libros sobreviven como copias en griego. Alrededor de 400 d. C., cuando la antigua civilización griega entró en declive, Arabia, China y la India tomaron de Europa la antorcha de la innovación matemática. Los estudiosos árabes tradujeron muchas de las obras griegas clásicas, y estas traducciones suelen ser nuestra principal fuente histórica para sus contenidos. El mundo árabe conocía la *Arithmetica*, y la desarrolló. Cuatro manuscritos árabes descubiertos en 1968 pueden ser traducciones de otros libros «perdidos» de la *Arithmetica*.

En algún momento próximo al final del siglo X el matemático persa al-Karaji planteó una pregunta que muy bien podría habérsele ocurrido a Diofanto. ¿Qué enteros pueden darse como diferencia común entre tres cuadrados racionales que forman una secuencia aritmética? Por ejemplo, los cuadrados enteros 1, 25 y 49 tienen una diferencia común 24. Es decir, 1 + 24 = 25 y 25 + 24 = 49. Al-Karaji vivió entre 953 y 1029 d. C. aproximadamente, de modo que quizá tuviera acceso a una copia de la *Arithmetica*, pero la primera traducción conocida fue hecha por Abu'l-Wafā en 998. Leonard Dickson, que escribió un resumen en tres volúmenes de la historia de la teoría de números, sugirió que el problema podría tener su

origen en algún momento antes de 972 en un manuscrito árabe anónimo.

En lenguaje algebraico el problema es: ¿para qué enteros d existe un número racional x tal que $x - d$, x y $x + d$ son cuadrados perfectos? Puede reenunciarse en una forma que es equivalente, aunque no sea obvia: ¿qué números enteros pueden ser el área de un triángulo rectángulo con lados racionales? Es decir: si a, b y c son racionales y $a^2 + b^2 = c^2$, ¿cuáles son los posibles valores enteros para $ab/2$? Los enteros que satisfacen estas condiciones equivalentes se llaman números congruentes. El término no está relacionado con otros usos de la palabra «congruente» en matemáticas, y eso lo hace algo confuso para un lector moderno. Sus orígenes se explican más adelante.

Algunos números no son congruentes: por ejemplo, puede demostrarse que 1, 2, 3 y 4 no son congruentes. Otros, tales como 5, 6 y 7 son congruentes. De hecho, el triángulo 3-4-5 tiene área $3\times4/2 = 6$, lo que prueba que 6 es congruente. Para demostrar que 7 es congruente, observemos que $(24/5)^2$, $(35/12)^2$ y $(337/60)^2$ tienen una diferencia común 7. Volveré a 5 en un momento. Procediendo caso por caso de esta manera se obtiene una larga lista de números congruentes, pero eso arroja poca luz sobre su naturaleza. Por muchos ejemplos que construyamos caso por caso, no podremos demostrar que un número entero concreto *no* es congruente. Durante siglos nadie sabía si 1 es congruente.

Ahora sabemos que el problema va mucho más allá de cualquier cosa que Diofanto pudiera haber resuelto. De hecho, esta pregunta

engañosamente simple todavía no ha sido respondida por completo. A lo más que hemos llegado es a una caracterización de números congruentes, descubierta por Jerrold Tunnell en 1983. La idea de Tunnell proporciona un algoritmo para decidir si un entero dado puede ocurrir o no contando sus representaciones como dos diferentes combinaciones de cuadrados. Con algo de ingenio este cálculo es factible para enteros muy grandes. La caracterización solo tiene un serio inconveniente: nunca se ha demostrado que sea correcta. Su validez depende de la solución de uno de los problemas del milenio, la conjetura de Birch-Swinnerton-Dyer. Esta conjetura proporciona un criterio para que una curva elíptica tenga solo un número finito de puntos racionales. Encontramos estas ecuaciones diofánticas en el capítulo 6 sobre la conjetura de Mordell y en el capítulo 7 sobre el último teorema de Fermat. Aquí vemos otra prueba de su papel prominente en las fronteras de la teoría de números.

La primera obra europea que hace referencia a estas cuestiones fue escrita por Leonardo de Pisa. Leonardo es más conocido por una secuencia de números extraños que él parece haber inventado, que aparecía en un problema aritmético sobre la progenie de algunos conejos muy poco realistas. Estos son los números de Fibonacci:

0 1 1 2 3 5 8 13 21 34 55 89...

en donde cada número, tras los dos primeros, es la suma de los dos anteriores. El padre de Leonardo era un oficial de aduanas llamado

Bonaccio, y el famoso sobrenombre significa «hijo de Bonaccio». No hay pruebas de que fuera utilizado durante la vida de Leonardo, y se cree que fue una invención del matemático francés Guillaume Libri en el siglo XIX[80]. Fuera como fuera, los números de Fibonacci tienen muchas propiedades fascinantes y son ampliamente conocidos. Incluso aparecen en la novela cripto-conspirativa *El código Da Vinci*, de Dan Brown.

Leonardo introdujo los números de Fibonacci en un libro de texto de aritmética, el *Liber abbaci* («Libro de cálculo») de 1202, cuyos objetivos principales eran llamar la atención europea sobre la nueva notación aritmética de los árabes, basada en los diez dígitos 0-9, y demostrar su utilidad. La idea ya había llegado a Europa a través del texto de al-Khwārizmī de 825 en su traducción latina *Algoritmi de numero indorum* («Sobre el cálculo con numerales hindúes»), pero el libro de Leonardo fue el primero escrito con la intención concreta de promover la adopción de la notación decimal en Europa. Buena parte del libro está dedicada a la aritmética práctica, especialmente al cambio de divisas. Pero Leonardo escribió otro libro, no tan bien conocido, que en muchos aspectos fue un sucesor europeo de la *Arithmetica* de Diofanto: su *Liber quadratorum* («Libro de cuadrados»).

Como Diofanto, él presentaba técnicas generales utilizando ejemplos especiales. Uno surgía de una pregunta de al-Karaji. En 1225 el emperador Federico II visitó Pisa. Él conocía la reputación matemática de Leonardo y parece haber decidido que sería divertido

[80] Leonardo Pisano Fibonacci, *The Book of Squares*, anotado y traducido por L. E. Sigler, Academia Press, 1987.

ponerle a prueba en un torneo matemático. Estos concursos públicos eran habituales en esa época. Los contendientes se planteaban preguntas unos a otros. El equipo del emperador estaba compuesto por Juan de Palermo y el maestro Teodoro. El equipo de Leonardo estaba compuesto por Leonardo. El equipo del emperador retó a Leonardo a encontrar un cuadrado que siga siendo un cuadrado cuando se le suma o se le resta 5. Como es habitual, los números deberían ser racionales. En otras palabras, ellos trataban de demostrar que 5 es un número congruente encontrando un racional concreto x para el que x - 5, x y x + 5 son cuadrados.
Esto no es en absoluto trivial; la solución más pequeña es:

$$x = \frac{1681}{144} = \left(\frac{41}{12}\right)^2$$

en cuyo caso

$$x-5 = \frac{961}{144} = \left(\frac{31}{12}\right)^2 \quad \text{y} \quad x+5 = \frac{2.401}{144} = \left(\frac{49}{12}\right)^2$$

Leonardo encontró una solución y la incluyó en el *Liber quadratorum*. Obtuvo la respuesta utilizando una fórmula general relacionada con la fórmula de Euclides/Diofanto para las tripletas pitagóricas. A partir de ella obtuvo tres cuadrados enteros con diferencia común 720, a saber, 31^2, 41^2 y 49^2. Luego dividió por 12^2 = 144 para obtener tres cuadrados con diferencia común 720/144,

que es 5.[81] En términos de tripletas pitagóricas, tomamos el triángulo 9, 40, 41 con área 180 y dividimos por 36 para obtener un triángulo con lados 20/3, 3/2, 41/6. Entonces su área es 5.

TABLA 3. Los diez primeros números congruentes y las correspondientes tripletas pitagóricas.

d	Tripleta pitagórica
5	3/2, 20/3, 41/6
6	3, 4, 5
7	24/5, 35/12, 337/60
13	780/323, 323/30, 106921/9690
14	8/3, 63/6, 65/6
15	15/2, 4, 17/2
20	3, 40/3, 41/3
21	7/2, 12, 25/2
22	33/35, 140/3, 4901/105
23	80155/20748, 41496/3485, 905141617/72306780

[81] Leonardo encontró una familia de soluciones.

$$\left(\frac{m^2+n^2}{2}\right)^2 - mn\left(m^2-n^2\right) = \left(\frac{m^2-2mn-n^2}{2}\right)^2$$

$$\left(\frac{m^2+n^2}{2}\right)^2 + mn\left(m^2-n^2\right) = \left(\frac{m^2+2mn-n^2}{2}\right)^2$$

donde m y n son impares. El papel de d lo desarrolla aquí el número $mn(m^2 - n^2)$, y x es $(m^2 + n^2)/2$. Escogiendo $m = 5$, $n = 4$ se llega a $mn(m^2 - n^2) = 180$. Además, $720 = 5 \times 6^2$. Dividiendo x por 6 se obtiene la respuesta.

Es en Leonardo donde encontramos la palabra latina *congruum* para un conjunto de tres cuadrados en secuencia aritmética. Más tarde Euler utilizó la palabra *congruere*, «venir juntos». En la Tabla 3 se da una lista de los diez primeros números congruentes y las correspondientes tripletas pitagóricas más sencillas. No hay pautas sencillas aparentes.

Muchos de los primeros progresos sobre esta cuestión fueron hechos por matemáticos islámicos, quienes demostraron que los números 5, 6, 14, 15, 21, 30, 34, 65, 70, 110, 154 y 190 son congruentes, junto con 18 números más grandes. A estos, Leonardo, Angelo Genocchi (1855) y André Gérardin (1915) añadieron 7, 22, 41, 69, 77 y otros 43 números menores que 1000. Leonardo afirmó en 1225 que 1 no es congruente, pero no dio ninguna demostración. En 1569 Fermat dio una. Para 1915 se habían determinado todos los números congruentes menores que 100, pero el problema encalló lentamente, y para 1980 el estatus de muchos números menores que 1000 seguía sin estar resuelto. La dificultad puede juzgarse por el descubrimiento de L. Bastien de que 101 es congruente. Los lados del correspondiente triángulo rectángulo son:

$$\frac{711024064578955010000}{118171431852779451900}$$

$$\frac{3967272806033495003922}{118171431852779451900}$$

$$\frac{4030484925899520003922}{118171431852779451900}$$

Él encontró estos números en 1914, a mano. Para 1986, ahora con ordenadores en escena, G. Kramarz había encontrado todos los números congruentes menores que 2000.

En algún momento se advirtió que una ecuación diferente pero relacionada

$$y^2 = x^3 - d^2x$$

tiene soluciones x, y en números naturales si y solo si d es congruente[82]. Esta observación es obvia en una dirección: el segundo miembro es el producto de x, $x - d$ y $x + d$, y si los tres son cuadrados, también lo es su producto. El recíproco también es bastante sencillo. Esta reformulación del problema lo coloca en el centro de un área rica y floreciente de la teoría de números. Para cualquier d dado esta ecuación hace y^2 igual a un polinomio cúbico

[82] Si $x - n$, x y $x + n$ son cuadrados, entonces también lo es su producto, que es $x^3 - n^2x$. Por consiguiente la ecuación $y^2 = x^3 - n^2x$ tiene una solución racional. Además, y no es cero, de lo contrario $x = n$ de modo que x y $2x$ son cuadrados, lo que es imposible puesto que $\sqrt{2}$ es irracional.
Recíprocamente, si x e y satisfacen la ecuación cúbica e y no es 0, entonces $a = (x^2 - n^2)/y$, $b = 2nx/y$, y $c = (x^2 + n^2)/y$ satisfacen las ecuaciones $a^2 + b^2 = c^2$ y $ab/2 = n$.

en *x*, y por consiguiente define una curva elíptica. De modo que el problema de los números congruentes es un caso especial de una pregunta que los teóricos de números desearían responder: ¿cuándo una curva elíptica tiene al menos un punto racional? Esta pregunta no es ni mucho menos sencilla, ni siquiera para el tipo especial de curva elíptica recién mencionado. Por ejemplo, 157 es un número congruente, pero el triángulo rectángulo *más simple* con dicha área tiene una hipotenusa

$$\frac{224403517704336969924557513090667486316094847204 1}{8912332268928859588025535178967163570016480830}$$

Antes de continuar tomaremos prestado el truco de Leonardo, el que le llevó de 720 a 5, y lo aplicaremos con toda generalidad. Si multiplicamos cualquier número congruente *d* por el cuadrado n^2 de un entero *n*, obtenemos también un número congruente. Simplemente tomamos una tripleta pitagórica racional correspondiente a un triángulo con área *d* y multiplicamos los números por *n*. Entonces el área del triángulo se multiplica por n^2. Lo mismo es cierto si dividimos los números por *n*; ahora el área se divide por n^2. Este proceso da un entero solo cuando el área tiene un factor cuadrado, de modo que cuando se buscan números congruentes es suficiente con trabajar con números que no tienen ningún factor cuadrado. Los primeros números sin factores cuadrados son:

1 2 3 5 6 7 10 11 13 14 15 17 19

Ahora podemos enunciar el criterio de Tunnell. Un número impar sin factores cuadrados d es congruente si y solo si el número de soluciones enteras x, y, z (positivas o negativas) de la ecuación

$$2x^2 + y^2 + 8z^2 = d$$

es exactamente el doble del número de soluciones de la ecuación

$$2x^2 + y^2 + 32z^2 = d$$

Un número par sin factores cuadrados d es congruente si y solo si el número de soluciones enteras x, y, z de la ecuación

$$8x^2 + 2y^2 + 16z^2 = d$$

es exactamente el doble del número de soluciones de la ecuación

$$8x^2 + 2y^2 + 64z^2 = d$$

Estos resultados son más útiles de lo que podría parecer a primera vista. Puesto que todos los coeficientes son positivos, los tamaños de x, y, z no pueden superar a ciertos múltiplos de la raíz cuadrada de d. Por lo tanto, el número de soluciones es finito y pueden encontrarse mediante una búsqueda sistemática, con algunos

atajos útiles. Estos son los cálculos completos para algunos ejemplos con d pequeño:

- Si $d = 1$ entonces las únicas soluciones de la primera ecuación son $x = 0$, $y = \pm 1$, $z = 0$. Lo mismo vale para la segunda ecuación. Por lo tanto, ambas ecuaciones tienen dos soluciones, y el criterio no es válido.
- Si $d = 2$ entonces las únicas soluciones de la primera ecuación son $x = \pm 1$, $y = 0$, $z = 0$. Lo mismo vale para la segunda ecuación. Por lo tanto, ambas ecuaciones tienen dos soluciones, y el criterio no es válido.
- Si $d = 3$ entonces las únicas soluciones de la primera ecuación son $x = \pm 1$, $y = \pm 1$, $z = 0$. Lo mismo vale para la segunda ecuación. Por lo tanto ambas ecuaciones tienen cuatro soluciones, y el criterio no es válido.
- Si $d = 5$ o 7 entonces la primera ecuación no tiene soluciones. Lo mismo vale para la segunda ecuación. Puesto que el doble de cero es cero, el criterio se satisface.
- Si $d = 6$ tenemos que utilizar el criterio para números pares. De nuevo ambas ecuaciones no tienen soluciones, y el criterio se satisface.

Estos cálculos sencillos muestran que 1, 2, 3, 4 ($= 2^{2\times 1}$) no son congruentes, pero 5, 6 y 7 sí lo son. El análisis puede ampliarse fácilmente, y en 2009 un equipo de matemáticos aplicó el test de Tunnell al primer billón de números y encontró exactamente 3.148.379.694 números congruentes. Los investigadores verificaron sus resultados realizando los cálculos dos veces, en diferentes

ordenadores y utilizando diferentes algoritmos escritos por dos grupos independientes. Bill Hart y Gonzalo Tornaria utilizaron el ordenador Selmer en la Universidad de Warwick. Mark Watkins, David Harvey y Robert Bradshaw utilizaron el ordenador Sage en la Universidad de Washington.

Sin embargo, hay una laguna en todos esos cálculos. Tunnell demostró que si un número d es congruente, entonces debe satisfacer su criterio. Por consiguiente, si el criterio falla, el número no es congruente. Esto implica, por ejemplo, que 1, 2, 3 y 4 no son congruentes. Sin embargo, él no pudo demostrar la inversa: si un número satisface su criterio, entonces debe ser congruente. Esto es lo que necesitamos para concluir que 5, 6 y 7 son congruentes. En estos casos particulares podemos encontrar tripletas pitagóricas apropiadas, pero eso no servirá en el caso general. Tunnell sí demostró que esta recíproca se sigue de la conjetura de Birch-Swinnerton-Dyer, pero esta sigue sin demostrar.

Como varios de los problemas del milenio, la conjetura de Birch-Swinnerton-Dyer es difícil incluso de enunciar. (¿Piensa usted que puede ganar un millón de dólares por hacer algo fácil? Puedo venderle un bonito puente, tirado de precio…). Sin embargo, la perseverancia recompensa, porque en el camino empezamos a apreciar las profundidades, y las largas tradiciones históricas, de la teoría de números. Si usted examina cuidadosamente el nombre de la conjetura verá que un guión es más largo que el otro. No es algo conjeturado por matemáticos llamados Birch, Swinnerton y Dyer, sino por Brian Birch y Peter Swinnerton-Dyer. Su enunciado

completo es técnico, pero trata de una cuestión básica en las ecuaciones diofánticas, es decir, ecuaciones algebraicas para las que buscamos soluciones en números naturales o racionales. La pregunta es simple: ¿cuándo tienen soluciones?

En el capítulo 6 sobre la conjetura de Mordell y en el capítulo 7 sobre el último teorema de Fermat encontramos algunos de los objetos más maravillosos en todas las matemáticas, las curvas elípticas. Mordell hizo lo que en ese momento era básicamente una conjetura aventurada, y supuso que el número de soluciones racionales de una ecuación algebraica en dos variables depende de la topología de la curva compleja asociada. Si el género es 0 —la curva es topológicamente una esfera— entonces las soluciones vienen dadas por una fórmula. Si el género es 1 —la curva es topológicamente un toro, que es equivalente a que sea una curva elíptica—, entonces todas las soluciones racionales pueden construirse a partir de una lista finita adecuada aplicando una estructura de grupo natural. Si el género es 2 o más —la curva es topológicamente un toro g-agujereado con $g \geq 2$—, entonces el número de soluciones es finito. Como vimos, Faltings demostró este notable teorema en 1983.

La propiedad más sorprendente de las soluciones racionales de ecuaciones de curva elíptica es que dichas soluciones forman un grupo, gracias a la construcción geométrica en la Figura 28 del capítulo 6. La estructura resultante se llama el grupo de Mordell-Weil de la curva, y a los teóricos de números les gustaría ser capaces de calcularlo. Eso incluye encontrar un sistema de

generadores: soluciones racionales a partir de las cuales pueden deducirse todas las demás utilizando repetidamente la operación del grupo. Si eso falla, nos gustaría al menos calcular algunas de las características básicas del grupo, tales como su tamaño. Aquí hay muchos detalles que aún no se comprenden. A veces el grupo es infinito, de modo que lleva a infinitas soluciones racionales; otras veces no lo es, y el número de soluciones racionales es finito. Sería útil poder decir cuál es cuál. De hecho, lo que realmente nos gustaría saber es la estructura abstracta del grupo.

La demostración de Mordell de que una lista finita genera todas las soluciones nos dice que el grupo debe estar construido a partir de un grupo finito y un grupo reticular. Un grupo reticular consiste en todas las listas de enteros de una longitud finita fija. Si la longitud es tres, por ejemplo, entonces el grupo consiste en todas las listas (m_1, m_2, m_3) de enteros, y las listas se suman de la manera obvia:

$$(m_1, m_2, m_3) + (n_1, n_2, n_3) = (m_1 + n_1, m_2 + n_2, m_3 + n_3)$$

La longitud de la lista se llama el rango del grupo (y geométricamente es la dimensión del retículo). Si el rango es 0, el grupo es finito. Si el rango es distinto de 0, el grupo es infinito. Así que para decidir cuántas soluciones hay no necesitamos la estructura completa del grupo. Todo lo que necesitamos es su rango. Y de eso es de lo que trata la conjetura de Birch-Swinnerton-Dyer.

En los años sesenta del siglo pasado, cuanto estaban naciendo los ordenadores, la Universidad de Cambridge tenía uno de los primeros, llamado EDSAC. Estas son las iniciales de Electronic Delay Storage Automatic Calculator (Calculator Automático de Almacenamiento Electrónico Diferido), y muestra lo orgullosos que estaban sus inventores por su sistema de memoria, que enviaba ondas sonoras a lo largo de tubos de mercurio y las redirigía de vuelta al principio. Tenía el tamaño de un camión grande, y recuerdo vívidamente cuando fue mostrado en 1963. Sus circuitos se basaban en miles de válvulas o tubos de vacío. Había enormes estantes a lo largo de las paredes con recambios para ser insertados cuando explotaba un tubo en la máquina, lo que sucedía muy a menudo.

Peter Swinnerton-Dyer estaba interesado en el lado diofántico de las curvas elípticas, y en particular quería entender cuántas soluciones habría si se reemplazara la curva por su análoga en un campo finito con un número primo p de elementos. Es decir, él quería estudiar el truco de Gauss de trabajar «módulo p». Utilizó el ordenador para calcular estos números para montones de primos y buscó pautas interesantes.

He aquí lo que él empezó a sospechar. Su supervisor, John William Scout («Ian») Cassels, era muy escéptico al principio, pero a medida que llegaban más y más datos empezó a creer en que podría haber algo en la idea. Lo que los experimentos en ordenador de Swinnerton-Dyer sugerían era esto. Los teóricos de números tienen un método estándar que reinterpreta cualquier ecuación en enteros

ordinarios en términos de enteros con algún módulo —recordemos la «aritmética de reloj» con módulo 12 en el capítulo 2—. Puesto que las reglas del álgebra se aplican en esta versión de la aritmética, cualquier solución de la ecuación original se convierte en una solución de la ecuación «reducida» con ese módulo. Puesto que los números implicados forman una lista finita —solo doce números en el caso de la aritmética de reloj, por ejemplo— se pueden encontrar todas las soluciones por ensayo y error. En particular, se pueden contar cuántas soluciones hay para cualquier módulo dado. Las soluciones con cualquier módulo también imponen condiciones sobre las soluciones enteras originales, y a veces pueden incluso demostrar que existen tales soluciones. Por eso, entre los teóricos de números existe el reflejo de reducir ecuaciones utilizando varios módulos, y los primos son una elección especialmente útil.

Por ello, para descubrir algo sobre una curva elíptica se pueden considerar todos los primos hasta un límite específico. Para cada primo se puede encontrar cuántos puntos se encuentran sobre la curva, módulo dicho primo. Birch advirtió que los experimentos en ordenador de Swinnerton-Dyer producen una pauta interesante si se divide el número de tales puntos por el primo en cuestión. Luego se multiplican todos estos factores, para todos los primos menores o iguales que uno dado, y se representan los resultados frente a primos sucesivos en papel logarítmico. Todos los datos parecen estar próximos a una línea recta cuya pendiente es el rango de la

curva elíptica. Esto llevó a una fórmula conjeturada para el número de soluciones asociadas a cualquier módulo primo[83].

No obstante, la fórmula no procede de la teoría de números: implica análisis complejo, el preferido del siglo XIX, que por algún milagro es mucho más elegante que el ya anticuado análisis real. En el capítulo 9 sobre la hipótesis de Riemann vimos cómo el análisis saca tentáculos en todas direcciones, que en particular tienen conexiones sorprendentes y poderosas con la teoría de números. La fórmula de Swinnerton-Dyer llevó a una conjetura más detallada sobre un tipo de función compleja que mencioné en el capítulo 9, llamada L-función de Dirichlet. Esta función es análoga, para las curvas elípticas, a la famosa función zeta de Riemann. Decididamente los dos matemáticos estaban tirando la casa por la ventana, porque en esa época no se sabía con seguridad que todas las curvas elípticas *tuvieran* L-funciones de Dirichlet. Era una conjetura aventurada con muy pocas pruebas a su favor. Pero conforme crecía el conocimiento del área, llegó a parecer cada vez más inspirada. No era un salto aventurado a lo desconocido: era un golpe de intuición matemática maravillosamente acertado y de gran alcance. En lugar de ponerse a hombros de gigantes, Birch y Swinnerton-Dyer se habían puesto sobre sus propios hombros — gigantes que podían cernirse en medio del aire.

---

[83] Esto es,

$$\prod_{p \le x} \frac{N_p}{p} \approx C(\log x)^r$$

donde r es el rango, C es una constante y ≈ significa que la razón de los dos miembros tiende a 1 cuando x tiende a infinito.

Una herramienta básica en análisis complejo es expresar una función utilizando una serie de potencias, similar a un polinomio pero con infinitos términos, que utiliza potencias cada vez mayores de la variable, que en esta área es tradicional llamarla *s*. Para descubrir lo que hace una función cerca de un punto concreto, digamos 1, se utilizan potencias de (*s* - 1). La conjetura de Birch-Swinnerton-Dyer afirma que si el desarrollo en serie de potencias cerca de 1 de una L-función de Dirichlet se parece a

$$L(C, s) = c(s - 1)^r + \text{términos de orden superior}$$

donde *c* es una constante distinta de cero, entonces el rango de la curva es *r*, y al revés. En el lenguaje del análisis complejo, esta afirmación toma la forma «$L(C, s)$ tiene un cero de orden r en $s = 1$».

El punto crucial aquí no es la expresión precisa requerida: es que dada cualquier curva elíptica, existe un cálculo analítico, que utiliza una función compleja relacionada, que nos dice de forma precisa cuántas soluciones racionales independientes tenemos que encontrar para especificarlas a todas.

Quizá la manera más sencilla de demostrar que la conjetura de Birch-Swinnerton-Dyer tiene contenido genuino es observar que el mayor rango conocido es 28. Es decir, existe una curva elíptica que tiene un conjunto de 28 soluciones racionales a partir de las cuales pueden deducirse todas las soluciones racionales. Además, ningún conjunto menor de soluciones racionales lo hace. Aunque se sabe que existen curvas de este rango, no se ha encontrado ningún

ejemplo explícito. El mayor rango conocido para un ejemplo explícito es 18. La curva, encontrada por Noam Elkies en 2006 es.

$$y^2 + xy = x^3 - 2617596009270588409631170178770120390355643896951 5x + 5106938147613148648974217710037377208977910325389056784832 6$$

Tal como está no es de la forma estándar «y^2 = cúbica en x», pero puede ponerse en dicha forma a expensas de hacer los números aún mayores. Se cree que el rango puede ser arbitrariamente grande, pero no se ha demostrado. Por lo que sabemos, el rango nunca puede superar un tamaño fijado.

Buena parte de lo que podemos demostrar concierne a curvas de rango 0 o 1. Cuando el rango es 0, hay un número finito de soluciones racionales. Cuando es 1, entonces una solución concreta lleva a casi todas las demás, con quizá un número finito de excepciones. Estos dos casos incluyen todas las curvas elípticas de la forma $y^2 = x^3 + px$ cuando p es un primo de la forma $8k + 5$ (tal como 13, 29, 37, y así sucesivamente). Se ha conjeturado que en estos casos el rango es siempre 1, lo que implica que hay infinitas soluciones racionales. Andrew Bremner y Cassels han demostrado que esto es cierto para todos estos primos hasta 100. Puede ser complicado encontrar soluciones que lleven a casi todas las demás, incluso cuando el rango es conocido y pequeño. Ellos encuentran que cuando p = 877 la solución *más simple* de este tipo es el número racional

$$\frac{375494528127162193105504069942092792346201}{6215987776871505425463220780697238044100}$$

Se han demostrado muchos teoremas relacionados con la conjetura de Birch-Swinnerton-Dyer, por lo general con hipótesis muy técnicas, pero los avances hacia una solución han sido relativamente escasos. En 1976 Coates y Wiles encontraron el primer indicio de que la conjetura podría ser cierta. Demostraron que un tipo especial de curva elíptica tiene rango 0 si la L-función de Dirichlet no se anula en 1. Para una curva elíptica semejante, el número de soluciones racionales de la ecuación diofántica es finito, quizá cero, y eso se puede deducir de la correspondiente L-función. Desde entonces ha habido algunos avances técnicos, aunque básicamente limitados a rangos 0 y 1. En 1990 Victor Kolyvagin demostró que la conjetura de Birch-Swinnerton-Dyer es cierta para rangos 0 y 1.

Conjeturas más detalladas, con mucho apoyo de ordenador, relacionan la constante *c* en la conjetura de Birch-Swinnerton-Dyer con varios conceptos en teoría de números. Hay análogos —igualmente enigmáticos— para campos de números algebraicos. También se sabe, en un sentido preciso, que la mayoría de las curvas elípticas tienen rango 0 o 1. En 2010 Manjul Bhargava y Arul Shankar anunciaron que habían demostrado que el rango promedio de una curva elíptica es a lo sumo 7/6. Si este y unos pocos más teoremas recientemente anunciados superan el escrutinio, la conjetura de Birch-Swinnerton-Dyer es cierta para

una proporción no nula de todas las curvas elípticas. Sin embargo, son las más simples y no representan en realidad a las curvas con una estructura más rica: rango 2 o más. Estas son un misterio casi total.

Capítulo 15

Ciclos complejos

La conjetura de Hodge

Algunas áreas de las matemáticas pueden estar relacionadas, de forma bastante directa, con sucesos e intereses cotidianos. No encontramos la ecuación de Navier-Stokes en nuestra cocina, pero todos entendemos lo que son los fluidos y tenemos una idea de cómo fluyen. Otras áreas pueden estar relacionadas con cuestiones esotéricas en la frontera de la ciencia: puede necesitarse un doctorado en física matemática para entender la teoría cuántica de campos, pero las analogías con la electricidad y el magnetismo, o imágenes intuitivas como «onda de probabilidad», llegan muy lejos. Algunas ideas pueden explicarse utilizando imágenes: la conjetura de Poincaré es un buen ejemplo. Pero otras desafían todos estos métodos de hacer accesibles conceptos abstractos difíciles.

La conjetura de Hodge, enunciada por el geómetra escocés William Hodge en 1950, es una de ellas. No es la demostración lo que plantea problemas, porque no la hay. Lo que causa problemas es el enunciado. He aquí el que figura en la página web del Instituto Clay, en una forma ligeramente modificada:

> *Sobre cualquier variedad algebraica compleja proyectiva y no singular, cualquier clase de Hodge es una combinación lineal racional de clases de ciclos algebraicos.*

Evidentemente tenemos trabajo por hacer. Las únicas palabras que tienen un sentido inmediato son «sobre, cualquiera, es, una» y «de». Otras son familiares: «variedad, clase, racional, ciclo». Pero las imágenes que evocan —elección en el supermercado, un aula con niños en la escuela, pensamiento no visceral, un dispositivo con ruedas y manillar— no son obviamente los significados que el Instituto Clay tiene en mente. El resto es, sin duda, propio de una jerga. Pero no es una jerga gratuita, con nombres complicados para cosas sencillas. Son nombres sencillos para cosas complicadas. No hay nombres ya disponibles para tales conceptos en el lenguaje ordinario, de modo que tomamos prestados algunos e inventamos otros.

Mirando el lado positivo, tenemos aquí una oportunidad real (como en «muchacho, tenemos oportunidades»). La conjetura de Hodge es presumiblemente más representativa de las matemáticas reales, las que hacen los matemáticos de los siglos XX y XXI, que cualquier otro tema en este libro. Enfocándola de la manera correcta obtenemos ideas valiosas sobre lo conceptualmente avanzada que es en realidad la frontera de las matemáticas. Comparada con las matemáticas de la escuela, es como el monte Everest frente a un grano de arena.

Entonces, ¿es simplemente un absurdo vano y pretencioso que se lleva a cabo en torres de marfil? Si ninguna persona ordinaria puede entender de qué se trata, ¿por qué alguien debería dedicar dinero de los impuestos a emplear a las personas que piensan en estas cosas? Permítame darle la vuelta. Supongamos que cualquier persona

pudiera entender todo en lo que piensan los matemáticos. ¿Sería usted feliz entonces pagando impuestos? ¿Hay que pagarles por sus conocimientos? Si todo fuera tan fácil y comprensible que tuviera un sentido inmediato para cualquier persona a la que preguntáramos al azar en la calle, ¿para qué tener matemáticos? Si todo el mundo supiera cómo manejar un soplete y soldar una juntura, ¿para qué tener fontaneros?

Yo no puedo mostrarle ninguna aplicación espectacular que se base en la conjetura de Hodge, pero puedo explicar su importancia dentro de las matemáticas. Las matemáticas modernas son un todo unificado, de modo que cualquier avance importante, en cualquier área nuclear, mostrará con el tiempo su valor en términos de euros y céntimos. Quizá no la encontremos hoy en nuestra cocina, pero mañana ¿quién sabe? Conceptos matemáticos íntimamente relacionados ya están mostrando su valor en varias áreas de la ciencia, que van desde la física cuántica y la teoría de cuerdas hasta los robots.

A veces las aplicaciones prácticas de nuevas matemáticas aparecen casi al instante. Otras veces tardan siglos. En este último caso, podría parecer más rentable esperar hasta que surja la necesidad de tales resultados y luego instaurar un programa de choque para desarrollarlos. Todos los problemas matemáticos que no tienen usos obvios e inmediatos deberían ser relegados a segundo plano hasta que los tengan. Sin embargo, si así lo hiciéramos iríamos siempre a remolque, pues las matemáticas pasaron algunos cientos de años persiguiendo las necesidades de la ciencia aplicada. Y podría no

estar nada claro qué idea necesitamos. ¿Estaría usted contento si nadie empezara a pensar en cómo hacer ladrillos hasta que usted contratara a un constructor para empezar a trabajar en una casa? Cuanto más original es un concepto matemático, menos probable es que saliera de un programa de choque.

Una mejor estrategia es dejar que algunas partes de las matemáticas se desarrollen según sus propias líneas y no esperar un rendimiento inmediato. No tratemos de seleccionar; permitamos que el edificio matemático crezca de forma orgánica. Los matemáticos son baratos: no necesitan un equipamiento costoso como los físicos de partículas (Gran Colisionador de Hadrones: 7500 millones de euros y contando). Se ganan la vida enseñando a estudiantes. Permitir que algunos de ellos dediquen parte de su tiempo a la conjetura de Hodge, si eso es lo que les gusta, es bastante razonable.

Voy a desmenuzar el enunciado de la conjetura de Hodge, palabra por palabra. El concepto más fácil es «variedad algebraica». Una consecuencia natural del uso de coordenadas por parte de Descartes es vincular la geometría y el álgebra (véase capítulo 3). Con su ayuda, el minúsculo juego de herramientas introducido por Euclides y sus sucesores —línea recta, círculo, elipse, parábola, hipérbola— se convirtió en un baúl sin fondo. Una línea recta, la base de la geometría euclídea, es el conjunto de puntos que satisfacen una ecuación algebraica apropiada, por ejemplo $y = 3x + 1$. Cambiemos 3 y 1 por otros números, y obtenemos otras rectas. Los círculos necesitan ecuaciones cuadráticas, como lo hacen las

elipses, parábolas e hipérbolas. En principio, cualquier cosa que se pueda enunciar geométricamente puede reformularse de manera algebraica, y al revés. Entonces, ¿las coordenadas hacen obsoleta la geometría? ¿Por qué utilizar dos herramientas cuando cada una de ellas hace el mismo trabajo que la otra?

En mi caja de herramientas en el garaje tengo un martillo y unas tenazas grandes. El trabajo del martillo es clavar clavos en la madera, el trabajo de las tenazas es volverlos a sacar. En principio, sin embargo, yo podría golpear los clavos utilizando las tenazas, y el martillo tiene una garra específicamente diseñada para extraer clavos. Entonces, ¿para que necesito las dos herramientas? Porque el martillo es mejor para algunas cosas y las tenazas son mejores para otras. Lo mismo pasa con la geometría y el álgebra: algunas maneras de pensar son más naturales utilizando la geometría, otras son más naturales utilizando el álgebra. Es el vínculo entre ellas lo que importa. Si el pensamiento geométrico se atasca, se pasa al álgebra. Si el pensamiento algebraico se atasca, se pasa a la geometría.

La geometría de coordenadas proporciona una nueva libertad para inventar curvas. Simplemente se escribe una ecuación y se miran sus soluciones. A menos que hayamos escogido una ecuación estúpida como $x = x$, deberíamos obtener una curva. (La ecuación $x = x$ tiene todo el plano como solución). Por ejemplo, yo podría escribir $x^3 + y^3 = 3xy$, cuyas soluciones están dibujadas en la Figura 45. Esta curva es el folio de Descartes, y no la encontraremos en

Euclides. El abanico de nuevas curvas que cualquiera puede inventar es literalmente infinito.

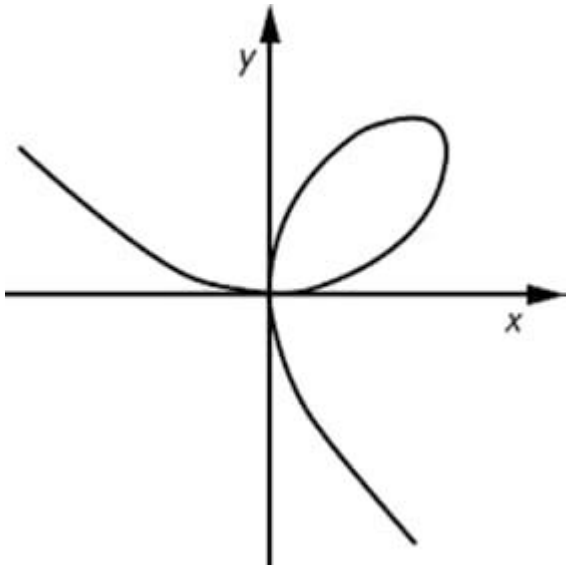


Figura 45. El folio de Descartes.

Un comportamiento reflejo entre matemáticos es generalizar. Una vez que alguien ha encontrado una idea interesante, podemos preguntar si sucede algo similar en un contexto más general. La idea de Descartes tiene al menos tres generalizaciones o modificaciones importantes, todas las cuales son necesarias para dar sentido a la conjetura de Hodge.

Primero, ¿qué sucede si trabajamos con espacios distintos del plano? El espacio euclídeo tridimensional tiene tres coordenadas (*x, y, z*) en lugar de dos. En el espacio, una ecuación define normalmente una superficie. Dos ecuaciones definen una curva, en donde se cruzan las correspondientes superficies. Tres ecuaciones determinan normalmente un punto. (Por «normalmente» quiero decir

que a veces puede haber excepciones, pero estas son muy poco habituales y satisfacen condiciones especiales. Vimos algo similar en el plano con la tonta ecuación $x = x$).

Una vez más, podemos definir nuevas superficies o curvas, que no se encuentran en Euclides, escribiendo nuevas ecuaciones. En el siglo XIX se puso de moda hacer eso. Uno podía hacer pública una nueva superficie si decía algo verdaderamente interesante sobre ella. Un ejemplo típico es el de una superficie introducida por Kummer en 1864, con la ecuación

$$x^4 + y^4 + z^4 - y^2z^2 - z^2x^2 - x^2y^2 - x^2 - y^2 - z^2 + 1 = 0$$

La Figura 46 muestra una imagen. Las propiedades de mayor interés son los 16 «puntos dobles» en donde la forma es como dos conos unidos punta con punta. Este es el máximo número posible para una superficie cuártica, cuya ecuación es de grado 4, y eso era suficientemente interesante para merecer su publicación.

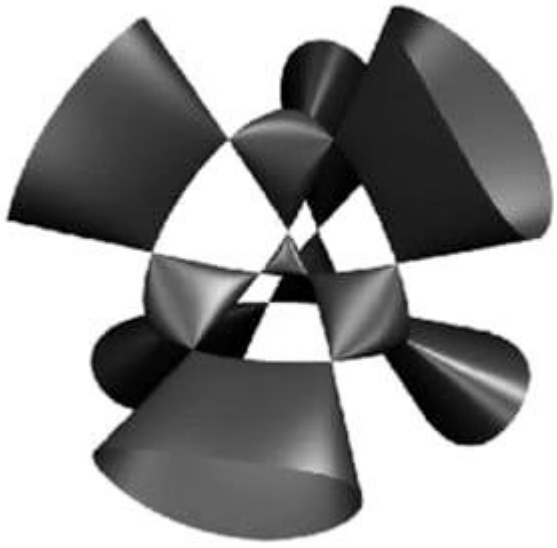


Figura 46. Superficie cuártica de Kummer con sus 16 puntos dobles.
© University College, Cork, Irlanda.

Para el siglo XIX los matemáticos habían experimentado los embriagadores deleites de los espacios de dimensiones más altas. No hay necesidad de pararse en tres coordenadas; ¿por qué no ensayar cuatro, cinco, seis..., un millón? Esto no es especulación ociosa. Es el álgebra de montones de ecuaciones en montones de variables, y esas se manifiestan en todo el paisaje matemático; por ejemplo, en el capítulo 5 sobre la conjetura de Kepler, y en el capítulo 8 sobre mecánica celeste. Tampoco era generalización ociosa: ser capaces de pensar estas cosas geométricamente, tanto como algebraicamente, es una herramienta poderosa que no debería estar restringida a espacios de dos o tres dimensiones, solo porque es en ellos donde se pueden dibujar imágenes y hacer modelos.

La palabra «dimensión» puede sonar impresionante y mística, pero en este contexto tiene un significado directo: cuántas coordenadas se necesitan. Por ejemplo, el espacio 4-dimensional tiene cuatro coordenadas (x, y, z, w), y por lo que concierne a las matemáticas, eso lo define. En cuatro dimensiones una sola ecuación define normalmente una «hipersuperficie» tridimensional, dos ecuaciones definen una superficie (dos dimensiones), tres ecuaciones definen una curva (una dimensión) y cuatro ecuaciones definen un punto (cero dimensiones). Cada nueva ecuación elimina una dimensión (una variable). Por eso podemos predecir que en el espacio de 17 dimensiones, once ecuaciones definen un objeto 6-dimensional,

excepto en los raros (y detectables) casos en donde algunas de las ecuaciones son superfluas.

Un objeto definido de esta manera se llama variedad algebraica. La palabra «variedad» aparece en lenguas como el francés y el español, y tiene un significado similar a «manifold» en inglés: básicamente, la palabra «many» (muchos). Por razones perdidas en las nieblas de la historia, «manifold» se asoció con la topología y la geometría diferencial —topología combinada con cálculo infinitesimal— mientras que «variedad» quedó asociada a la geometría algebraica[84]. Utilizar nombres diferentes evita la confusión, de modo que ambos cuajaron. Una variedad algebraica podría haberse llamado un «espacio multidimensional definido por un sistema de ecuaciones algebraicas», pero es fácil ver por qué nadie lo hizo[iv].

Una segunda manera atractiva de generalizar las nociones de geometría de coordenadas es permitir que las coordenadas sean números complejos. Recordemos que el sistema de los números complejos incluye un nuevo tipo de número, i, cuyo cuadrado es -1. ¿Por qué complicarlo todo de esa manera? Porque las ecuaciones algebraicas se comportan mucho mejor en el campo complejo. En el campo real una ecuación cuadrática puede tener dos soluciones o ninguna. (También puede tener solo una, pero en cierto sentido es la misma solución que aparece dos veces). En el campo complejo una ecuación cuadrática tiene siempre dos soluciones (de nuevo contando las multiplicidades correctamente). Para algunos fines, esta es una propiedad mucho más agradable. Podemos decir

[84] La razón más probable es que estas son las traducciones naturales de las lenguas utilizadas por los matemáticos más importantes en las dos áreas.

«resolver la ecuación para la séptima variable» y tener la seguridad de que tal solución existe realmente.

Por agradable que pueda ser a este respecto, la geometría algebraica compleja tiene características que hacen necesario un tiempo para habituarse a ella. Con variables reales, una recta puede cortar a un círculo, o ser tangente al mismo, o no cortarlo en absoluto. Con variables complejas, la tercera opción desaparece. No obstante, una vez que uno se ha acostumbrado a estos cambios, las variedades algebraicas complejas se comportan mucho mejor que las reales. A veces las variables reales son esenciales, pero para la mayoría de los fines el campo complejo es una mejor elección. En cualquier caso, ahora sabemos qué es una variedad algebraica compleja.

¿Qué pasa con «proyectiva»? Esta es la tercera generalización y requiere una noción de espacio ligeramente diferente. La geometría proyectiva surgió del interés en la perspectiva por parte de los pintores del Renacimiento y elimina el comportamiento excepcional de las rectas paralelas. En la geometría de Euclides dos rectas o bien se cortan o bien son paralelas: no se encuentran por mucho que se prolonguen. Imagínese ahora situado en un plano infinito, pincel en mano, caballete montado, caja de pinturas preparada, con un par de rectas paralelas que apuntan hacia la lejana puesta de sol como vías de tren infinitamente largas. ¿Qué es lo que usted ve, y qué dibujaría? No dos líneas que no se encuentran. En su lugar, las líneas parecen converger para encontrarse en el horizonte.

¿A qué parte del plano corresponde el horizonte? Es la parte en donde se encuentran las paralelas. Pero no hay tal cosa. El

horizonte es la frontera, en su cuadro, de la imagen del plano. Si todo va bien, eso debería ser sin duda la imagen de la frontera del plano. Pero un plano no tiene frontera; continúa indefinidamente. Todo esto es un poco lioso. Es como si faltara una parte del plano euclídeo. Si se «proyecta» un plano (el plano con las vías de tren) sobre otro plano (el lienzo en el caballete) se obtiene una línea en la imagen, el horizonte, que no es la proyección de ninguna línea en el plano.

Hay una manera de deshacerse de esta intrigante anomalía: añadir al plano euclídeo una denominada recta en el infinito que representa el horizonte que falta. Ahora todo se hace mucho más sencillo. Dos líneas siempre se cortan en un punto; la vieja noción de líneas paralelas corresponde al caso en que las dos líneas se encuentran en el infinito. Esta idea, adecuadamente interpretada, puede convertirse en matemáticas perfectamente razonables. El resultado se llama geometría proyectiva. Es una disciplina muy elegante, y los matemáticos de los siglos XVIII y XIX la amaban. Con el tiempo agotó lo que tenía que decir, hasta que los matemáticos del siglo XX decidieron generalizar la geometría algebraica a espacios multidimensionales y utilizar números complejos. En ese momento se hizo claro que también podríamos tirar la casa por la ventana y estudiar soluciones complejas de ecuaciones algebraicas en el espacio proyectivo antes que soluciones reales en el espacio euclídeo.

Permítame resumir. Una variedad algebraica compleja proyectiva es como una curva, definida por una ecuación algebraica, pero:

- El número de ecuaciones y variables puede ser cualquiera que queramos (variedad algebraica).
- Las variables pueden ser complejas en lugar de reales (compleja).
- Las variables pueden tomar valores infinitos en una forma razonable (proyectiva).

Para concluir con esto, hay otro término con el que se puede tratar fácilmente: no singular. Significa que la variedad es suave, sin cordilleras abruptas o lugares donde la forma es más complicada que una región suave de espacio. La superficie de Kummer es singular en esos 16 puntos dobles. Por supuesto, tenemos que explicar qué significa «suave» cuando las variables son complejas y algunas pueden valer infinito, pero eso es una técnica rutinaria.

Estamos casi a mitad de camino en el enunciado de la conjetura de Hodge. Sabemos de qué estamos hablando, pero no sabemos cómo pensaba Hodge que debía de comportarse. Ahora tenemos que abordar los aspectos más profundos y más técnicos: ciclos algebraicos, clases, y (especialmente) clases de Hodge. Sin embargo, puedo revelar su esencia en pocas palabras. Son artificios técnicos que proporcionan una respuesta parcial a una pregunta muy básica sobre nuestra superficie generalizada: *¿qué forma tiene?* Los únicos términos restantes, «combinación lineal racional», proporcionan lo que todo el mundo espera que sea la respuesta correcta a esa pregunta.

Veamos hasta dónde hemos llegado. Ya entendemos qué tipo de enunciado es la conjetura de Hodge. Nos dice que dada una

superficie generalizada definida por ciertas ecuaciones, se puede calcular qué forma tiene haciendo algo de álgebra con cosas llamadas ciclos. Yo podría haberlo dicho en la primera página de este capítulo, pero en esa etapa no habría tenido más sentido que el enunciado formal. Ahora que sabemos lo que es una variedad, todo empieza a encajar.

También empieza a sonar a topología. «Encontrar la forma haciendo cálculos algebraicos» recuerda sorprendentemente las ideas de Poincaré sobre invariantes algebraicos para espacios topológicos. Así que el último paso requiere una discusión de la topología algebraica. Entre los descubrimientos de Poincaré había tres tipos importantes de invariantes, definidos en términos de tres conceptos: homotopía, homología y cohomología. El que nos interesa es la cohomología, que, por supuesto, por si no lo supiera, es el más difícil de explicar.

Pienso que simplemente tenemos que saltarlo.

En el espacio tridimensional con coordenadas reales, una esfera y un plano se cortan (si es que lo hacen) en un círculo. La esfera es una variedad (omitiré el adjetivo «algebraica» cuando hablemos de variedades), el círculo es una variedad, y el círculo está contenido en la esfera. Le llamamos una *subvariedad*. Con más generalidad, si tomamos las ecuaciones (muchas variables, complejas, proyectivas) que definen una variedad, y añadimos otras ecuaciones, entonces normalmente perdemos algunas de las soluciones: aquellas que no satisfacen las nuevas ecuaciones. Cuantas más ecuaciones tengamos, menor se hace la variedad. El sistema de ecuaciones

ampliado define alguna parte de la variedad original, y esta parte es una variedad por sí misma: una subvariedad.

Cuando contamos el número de soluciones de una ecuación polinómica puede ser conveniente contar el mismo punto más de una vez. Desde esta perspectiva, el conjunto de soluciones consiste en un número de puntos a cada uno de los cuales «asociamos» un número, su multiplicidad. Podríamos, por ejemplo, tener las soluciones 0, 1 y 2 con multiplicidades 3, 7 y 4, respectivamente. El polinomio sería entonces $x^3(x - 1)^7 (x - 2)^4$, si quiere saberlo. Cada uno de los tres puntos $x = 0$, 1 o 2 es una subvariedad (bastante trivial) de los números complejos. De modo que las soluciones de este polinomio pueden describirse como una lista de tres subvariedades, con un número entero asociado a cada una como una etiqueta.

Un ciclo algebraico es similar. En lugar de puntos simples utilizamos cualquier lista finita de subvariedades. A cada una de ellas podemos asignar una etiqueta numérica que no tiene por qué ser un número natural. Podría ser un entero negativo, podría ser un número racional, podría ser real e incluso un número complejo. Por varias razones, la conjetura de Hodge utiliza números racionales como etiquetas. A esto es a lo que se refiere «combinación lineal racional». Así, por ejemplo, nuestra variedad original podría ser la esfera unidad en un espacio 11-dimensional, y esta lista podría tener este aspecto:

Una hiperesfera 7-dimensional (con ecuaciones tales y cuales) con etiqueta 22/7.

Un toro (con ecuaciones tales y cuales) con etiqueta -4/5.
Una curva (con ecuaciones tales y cuales) con etiqueta 413/6.
No trate de representarlo, o si lo hace, piense como un dibujante de comics: tres globos garabateados con pequeñas etiquetas. Cada uno de estos dibujos, cada lista, constituye un ciclo algebraico.
¿Por qué tanto revuelo y molestia para inventar algo tan abstracto? Porque capta aspectos esenciales de la variedad algebraica original. Los geómetras algebraicos están tomando prestado un truco de los topólogos.
En el capítulo 10 sobre la conjetura de Poincaré consideramos una hormiga cuyo universo es una superficie. ¿Cómo puede la hormiga calcular qué forma tiene su universo cuando no puede salir fuera y echar una mirada? En particular, ¿cómo puede distinguir una esfera de un toro? La solución presentada allí implicaba lazos cerrados, trayectos de autobús topológicos. La hormiga sigue estos lazos, descubre lo que sucede cuando se unen y calcula un invariante algebraico del espacio llamado grupo fundamental. «Invariante» significa que espacios topológicamente equivalentes tienen el mismo grupo fundamental. Si los grupos son diferentes, también lo son los espacios. Este es el invariante que llevó a Poincaré a su conjetura. Sin embargo, no es fácil para la pobre hormiga examinar todos los lazos posibles en su universo, y este comentario refleja genuinas sutilezas matemáticas en el cálculo del grupo fundamental. Existe un invariante más práctico, y Poincaré también lo investigó. Deambular por lazos se llama homotopía. Esta alternativa tiene un nombre parecido: homología.

Voy a mostrarle la versión más simple y más concreta de homología. Los topólogos mejoraron rápidamente esta versión, la agilizaron, la generalizaron y la transformaron en una enorme máquina matemática llamada álgebra homológica. Esta versión simple da una idea más escueta de cómo es el tema, pero es todo lo que necesitamos.

La hormiga empieza inspeccionando su universo para hacer un mapa. Como un topógrafo humano, cubre su universo con una red de triángulos. La condición crucial es que ningún triángulo debería rodear a un agujero en la superficie, y la manera de asegurarlo es crear los triángulos pegando parches de goma en la superficie, como alguien que repara un pinchazo en una rueda de bicicleta. Entonces cada triángulo tiene un interior bien definido que es topológicamente igual al interior de un triángulo ordinario en el plano. Los topólogos llaman a un parche semejante un disco topológico, porque también es equivalente a un círculo y su interior. Para ver por qué, examinemos la Figura 36 del capítulo 10, donde un triángulo es deformado de forma continua hasta que se hace un círculo. No es posible ajustar un parche de este tipo a un triángulo que rodea a un agujero, porque el agujero crea un túnel que enlaza el interior del triángulo con su exterior. El parche tendría que dejar la superficie, y a la hormiga no se le permite hacer eso.

Ahora la hormiga ha creado una *triangulación* de su universo. La condición sobre los parches asegura que la topología de la superficie —su forma, en el sentido de equivalencia topológica— puede reconstruirse si todo lo que se conoce es la lista de triángulos, junto

con qué triángulos son adyacentes a cuáles otros. Si fuéramos a Ikea y compráramos un Universo de Hormiga ensamblable con triángulos adecuadamente etiquetados y luego uniéramos el borde A con el borde AA, el borde B con el borde BB, y así sucesivamente, podríamos construir la superficie. La hormiga está confinada en la superficie, de modo que no puede hacer un modelo, pero puede estar segura de que en principio su mapa contiene la información que necesita. Para extraer dicha información tiene que realizar un cálculo. Cuando lo hace, la hormiga ya no tiene que contemplar la infinitud de todos los lazos posibles, pero tiene que contemplar una gran cantidad de ellos: todos los lazos cerrados que recorren los bordes de su red elegida.

En homotopía preguntamos si un lazo dado puede contraerse de forma continua hasta un punto. En homología hacemos una pregunta diferente: ¿forma el lazo la frontera de un disco topológico? Es decir, ¿se pueden ajustar uno o más parches triangulares de modo que el resultado sea una región sin agujeros y la frontera de dicha región sea el lazo concernido?

La Figura 47 (*izquierda*) muestra parte de una triangulación de una esfera, un lazo cerrado y el disco topológico del cual es frontera. Con las técnicas correctas puede demostrarse que *cualquier* lazo en una triangulación de una esfera es una frontera: los parches triangulares y, más en general, los discos topológicos, son detectores de agujeros, e intuitivamente una esfera no tiene agujeros. Sin embargo, un toro sí tiene un agujero, y de hecho algunos lazos en un toro no son fronteras. La Figura 47 (*derecha*)

muestra un lazo semejante, que atraviesa el agujero central. En otras palabras: recorriendo una lista de lazos y descubriendo cuáles de ellos son fronteras, la hormiga puede distinguir un universo esférico de un universo toroidal.

Si la hormiga es tan inteligente como Poincaré y los demás topólogos de su tiempo, puede transformar esta idea en un elegante invariante topológico, el grupo de homología de su superficie. La idea básica es «sumar» dos lazos dibujando ambos. Sin embargo, eso no es un lazo, de modo que tenemos que volver al principio y empezar de nuevo. Al mismo principio, de hecho; vuelta a los días en que nos introdujimos por primera vez en el álgebra. Mi profesor de matemáticas empezaba señalando que se puede sumar un número de manzanas a un número de manzanas y obtener un número total de manzanas. Pero no se pueden sumar manzanas y naranjas a menos que se cuente todo como una fruta.

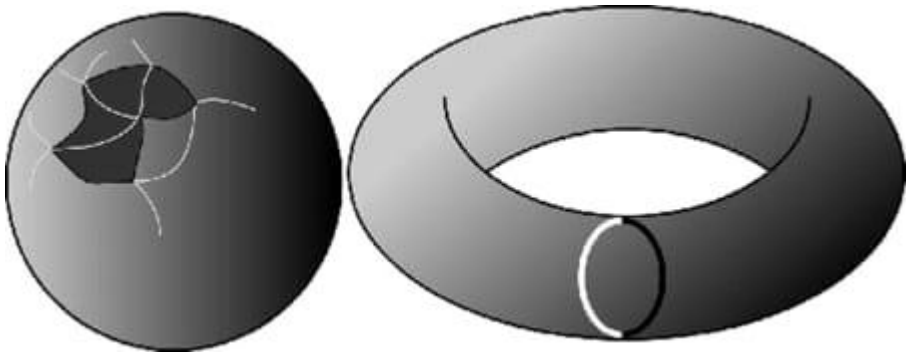


Figura 47. Izquierda: Parte de una triangulación de una esfera, un lazo cerrado (líneas negras), y el disco del que es frontera (sombreado oscuro). Derecha: Lazo en un toro que no es la frontera de un disco (la parte más clara está detrás).

Esto es cierto en aritmética, incluso si allí hay que tener cuidado en no utilizar la misma manzana dos veces, pero no es cierto en álgebra. Allí se pueden sumar manzanas con naranjas, aunque se mantengan distintas. De hecho, en matemáticas avanzadas es un tópico sumar cosas que cabría pensar que nadie en su sano juicio hubiera inventado, y mucho menos quisiera sumar. La libertad para hacer cosas de este tipo resulta ser sorprendentemente útil e importante, y los matemáticos que lo hacían no estaban locos después de todo; al menos, no en ese aspecto.

Para entender algunas de las ideas que reúne la conjetura de Hodge tenemos que poder sumar manzanas y naranjas sin amontonarlas como simples frutas. La manera de sumarlas no es realmente muy difícil. Lo que es difícil es aceptar que hay un sentido en hacerlo. Muchos de nosotros ya hemos encontrado una versión de este potencial obstáculo conceptual. Mi profesor contaba a la clase que las letras representaban números desconocidos, con diferentes letras para diferentes incógnitas. Si tuviéramos a manzanas y otras a manzanas, el número total de manzanas sería $a + a = 2a$. Y eso funcionaba cualquiera que pudiera ser el número de manzanas. Si tuviéramos $3a$ manzanas y sumáramos $2a$ manzanas, el resultado sería $5a$, cualquiera que pudiera ser el número de manzanas. El símbolo, y lo que representaba, no importaba: si tuviéramos $3b$ naranjas y sumáramos $2b$ naranjas, el resultado sería $5b$[85]. Pero

[85] No estoy seguro de por qué b no era el número de bananas. ¿Quizá porque en la Gran Bretaña de la posguerra las bananas eran artículos exóticos que apenas se veían en las tiendas?

¿qué sucedía cuando teníamos $3a$ manzanas y $2b$ naranjas? ¿Qué era $3a + 2b$?

$$3a + 2b$$

Eso era. No podíamos simplificar la suma y hacerla 5 algos —al menos no sin algunas manipulaciones que incluían una nueva categoría, fruta, y algunas nuevas ecuaciones. Eso era lo más que podíamos hacer: vivir con ello. Sin embargo, una vez que se daba ese paso, se podían hacer sumas como:

$$(3a + 2b) + (5a - b) = 8a + 4b$$

sin introducir ninguna idea nueva. O nuevos tipos de fruta.

Había algunas reservas. Ya he advertido que si se suma una manzana a una manzana, solo obtenemos dos manzanas si la segunda manzana es diferente de la primera. Lo mismo sucede con combinaciones más complicadas de manzanas y naranjas. El álgebra supone que con el fin de hacer las sumas, todas las manzanas implicadas son diferentes. De hecho, suele ser razonable hacer esta hipótesis, incluso en casos donde dos manzanas —o cualquier otra cosa que estemos sumando— podrían ser en realidad la misma. Una manzana más la misma manzana es una manzana con multiplicidad dos.

Una vez que nos acostumbramos a esta idea, podemos utilizarla para cualquier cosa. Un cerdo más el mismo cerdo es ese cerdo con

multiplicidad dos: cerdo + cerdo = 2 cerdo, cualquier cosa que sea cerdo. Un cerdo más una vaca es cerdo + vaca. Un triángulo más tres círculos es triángulo + 3 círculo. Una superduperesfera más tres quasimontones hiperelípticos es

superduperesfera + 3 cuasimontonhiperelíptico

cualquier cosa que esa jerga signifique (que, aquí, es nada).

Incluso podemos permitir números negativos y hablar de tres cerdos menos once vacas: 3 cerdo - 11 vaca. No tengo la menor idea de lo que sean menos once vacas, pero puedo confiar en que si a eso le sumo seis vacas, he obtenido -5 vacas[86]. Es un juego formal que se juega con símbolos, y no se necesita ninguna interpretación más realista, útil o —a menudo— posible. Podríamos admitir números reales: π cerdos menos √2 vacas. O números complejos. O cualquier tipo de número fantástico que haya inventado o invente en el futuro cualquier matemático. La idea puede hacerse algo más respetable si consideramos los números como *etiquetas* unidas a los cerdos y las vacas. Ahora π cerdos menos √2 vacas pueden considerarse como un cerdo etiquetado π junto con una vaca etiquetada √-2. La aritmética se aplica a las etiquetas, no a los animales.

La conjetura de Hodge implica una construcción de este tipo, con algunos detalles extra. En lugar de animales, utiliza curvas,

[86] Este es un chiste matemático estándar: un biólogo, un estadístico y un matemático están sentados en la terraza de un café observando a la gente que pasa. Un hombre y una mujer entran en un edificio al otro lado de la calle. Diez minutos más tarde salen acompañados de un niño. «Ellos se han reproducido», dice el biólogo. «No —dice el estadístico—, es un error observacional. En promedio, dos personas y media fueron en cada dirección». «No, no, no —dice el matemático—. Es perfectamente obvio. Si alguien entra ahora, el edificio estará vacío».

superficies y sus análogas en más altas dimensiones. Por extraño que pueda parecer, el resultado no es un absurdo abstracto sino una conexión profunda entre topología, álgebra, geometría y análisis.

Para establecer el formalismo de homología queremos sumar lazos, pero no de la manera en que lo hicimos en el caso del grupo fundamental. En su lugar, lo hacemos de la manera que me contaba mi profesor. Simplemente escribimos los lazos y ponemos un signo + entre ellos. Para darle sentido trabajamos no con lazos únicos sino con conjuntos finitos de ellos. Etiquetamos cada lazo con un entero que cuenta cuántas veces ocurre. Llamamos un *ciclo* a dicho conjunto etiquetado. Ahora la hormiga puede sumar dos ciclos cualesquiera agrupándolos y sumando las etiquetas correspondientes, y el resultado es otro ciclo. Quizá debería haber utilizado bicicletas, no autobuses, en mi imagen para los viajes de la hormiga en el capítulo 10.

Cuando estábamos construyendo el grupo fundamental, donde la «suma» une lazos extremo con extremo, había una pega técnica. Sumar el lazo trivial a un lazo no daba *exactamente* el mismo lazo, de modo que el lazo cero se comportaba mal. Sumar un lazo a su inverso no daba exactamente el lazo trivial, de modo que los inversos no se comportaban correctamente. La vía de escape era considerar que los lazos son el mismo si uno pudiera deformarse hasta el otro.

En el caso de la homología, este no es el problema. Hay un ciclo cero (todas las etiquetas cero), y todo ciclo tiene un inverso

(convierte toda etiqueta en su negativo), de modo que obtenemos un grupo. El problema es que se trata del grupo equivocado. No nos dice nada sobre la topología del espacio. Para solucionarlo utilizamos un truco similar, y adoptamos una visión más laxa sobre qué ciclos deberían contar como cero. La hormiga divide el espacio en parches triangulares, y la frontera de cada parche es topológicamente bastante trivial: podemos contraerla hasta un punto empujándola hacia el centro de su parche. De modo que exigimos que estos ciclos frontera sean equivalentes al ciclo cero. Es un poco como convertir números ordinarios en aritmética de reloj fingiendo que el número 12 es irrelevante, de modo que puede hacerse cero. Aquí convertimos ciclos en homología fingiendo que cualquier ciclo frontera es irrelevante.

Las consecuencias de este fingimiento son espectaculares. Ahora el álgebra de ciclos está afectada por la topología del espacio. El grupo de ciclos módulo fronteras es un útil invariante topológico, el grupo de homología de la superficie. A primera vista depende de qué triangulación escoja la hormiga, pero como en el caso de la característica de Euler, diferentes triangulaciones de la misma superficie llevan al mismo grupo de homología. Así que la hormiga ha inventado un invariante algebraico que puede distinguir superficies diferentes. Es un poco enrevesado, pero nunca se obtienen buenos invariantes sin hacer un trabajo duro en alguna parte del camino. Este es tan efectivo que puede distinguir no solo la esfera del toro sino también un toro 2-agujereado de un toro de 5-agujereado, y lo mismo para cualquier otro número de agujeros.

La homología puede parecer un trabalenguas, pero abrió una rica vena de invariantes topológicos y se basa en simples ideas geométricas: lazos, fronteras, agrupar conjuntos, hacer aritmética con etiquetas. Considerando que la pobre hormiga está confinada en su superficie, es sorprendente que la criatura pueda descubrir algo importante sobre la forma de su universo solo con pegar parches triangulares, hacer un mapa y utilizar un poco de álgebra.

Hay una manera natural de extender la homología a dimensiones más altas. El análogo 3-dimensional de un triángulo es un tetraedro; tiene 4 vértices, 6 aristas, 4 caras triangulares y una única «cara» 3-dimensional, su interior. En general, en n dimensiones podemos definir un n-simplex con $n + 1$ vértices, unidos en pares por todas las aristas posibles, que a su vez forman triángulos que se ensamblan para crear tetraedros, y así sucesivamente. Ahora es fácil definir ciclos, fronteras y homología, y de nuevo podemos construir un grupo sumando (clases de homología de) ciclos. De hecho, ahora obtenemos toda una serie de grupos: uno para ciclos 0-dimensionales (puntos), otro para ciclos 1-dimensionales (líneas), otro para ciclos 2-dimensionales (triángulos), y así sucesivamente, hasta la dimensión del propio espacio. Estos son el 0-ésimo, primero, segundo, y así sucesivamente, grupos de homología del espacio. Hablando en términos generales, hacen precisa la noción de agujeros, de varias dimensiones, en el espacio: ¿existen, cuántos hay y cómo se relacionan entre sí?

Eso, entonces, es la homología, y es casi lo que necesitamos para entender lo que *dice* la conjetura de Hodge. Sin embargo, lo que necesitamos en realidad es un concepto íntimamente relacionado llamado *cohomología*. En 1893 Poincaré advirtió una curiosa coincidencia en la homología de cualquier variedad: la lista de grupos de homología se lee igual al revés. Para una variedad de dimensión 5, pongamos por caso, el 0-ésimo grupo de homología es el mismo que el 5.º, el 1.er grupo de homología es el mismo que el 4.º, y el 2.º grupo de homología es el mismo que el 3.º. Él se dio cuenta de que esto no podía ser solo una coincidencia y la explicó en términos del dual de una triangulación, que encontramos en el capítulo 4 en conexión con los mapas. Este es una segunda triangulación en la que cada triángulo se reemplaza por un vértice, cada arista entre dos triángulos por una arista que enlaza los correspondientes nuevos vértices, y cada punto por un triángulo, como en la Figura 9 del capítulo 4. Nótese cómo la dimensión aparece en orden inverso: triángulos 2-dimensionales se convierten en puntos 0-dimensionales, y recíprocamente; aristas 1-dimensionales siguen siendo aristas 1-dimensionales porque 1 está en el centro.

Resulta útil distinguir las dos listas, incluso si dan los mismos invariantes. Cuando se generaliza el sistema entero y se formula en términos abstractos, las triangulaciones desaparecen y la triangulación dual ya no tiene sentido. Lo que sobrevive son dos series de invariantes topológicos, llamados grupos de homología y grupos de cohomología. Todo concepto en homología tiene un dual,

cuyo nombre se forma normalmente añadiendo «co» delante. Así, en lugar de ciclos tenemos cociclos, y en lugar de que dos ciclos sean homólogos tenemos dos cociclos que son cohomólogos. Las clases mencionadas en la conjetura de Hodge son clases de cohomología, y estas son colecciones de cociclos que son cohomólogos entre sí.

Homología y cohomología no nos dicen todo lo que nos gustaría saber sobre la forma de un espacio topológico —distintos espacios pueden tener la misma homología y cohomología— pero proporcionan mucha información útil y un marco sistemático en el que calcularla y utilizarla.

Una variedad algebraica, ya sea real, compleja, proyectiva o no, es un espacio topológico. Por consiguiente tiene una forma. Para descubrir cosas útiles sobre la forma, pensamos como topólogos y calculamos los grupos de homología y cohomología. Pero los ingredientes naturales en geometría algebraica no son objetos geométricos como triangulaciones y ciclos. Son los objetos que más fácilmente podemos describir por ecuaciones algebraicas. Volvamos atrás y examinemos la ecuación para la superficie de Kummer. ¿Cómo se relacionaría con una triangulación? No hay nada en la fórmula que sugiera triángulos.

Quizá necesitemos empezar de nuevo. En lugar de triángulos deberíamos utilizar los bloques constituyentes básicos de las variedades, que son subvariedades definidas imponiendo ecuaciones extra. Ahora tenemos que redefinir los ciclos: en lugar de conjuntos de triángulos con etiquetas enteras, utilizamos conjuntos de subvariedades con cualesquiera etiquetas que mejor sirvan. Por

diversas razones —sobre todo porque la conjetura de Hodge es falsa si utilizamos etiquetas enteras— los números racionales son la elección razonable. La pregunta de Hodge se reduce a esto: ¿recoge esta nueva definición de homología y cohomología todo lo que recoge la definición topológica? Si su conjetura es cierta, entonces la herramienta del ciclo algebraico es suficientemente aguda para igualar al cincel cohomológico de la topología. Si es falsa, entonces el ciclo algebraico es un instrumento romo.

Excepto… lo siento, he puesto demasiados huevos en el pudin. La conjetura dice que basta con utilizar un *tipo* particular de ciclo algebraico, uno que vive en una clase de Hodge. Para explicarlo necesitamos otro ingrediente en una ya rica mezcla: el análisis. Uno de los conceptos más importantes en el análisis es el de una ecuación diferencial, que es una condición sobre los ritmos a que cambian las variables (véase capítulo 8). Casi toda la física matemática de los siglos XVIII, XIX y XX modela la naturaleza utilizando ecuaciones diferenciales, e incluso en el siglo XXI la mayor parte lo hace. En la década de 1930 esta idea llevó a Hodge a un nuevo cuerpo de técnica, ahora llamado teoría de Hodge. Se relaciona de forma natural con muchos otros métodos potentes en el área general del análisis y la topología.

La idea de Hodge era utilizar una ecuación diferencial para organizar las clases de cohomología en tipos característicos. Cada pieza tiene estructura extra, que puede utilizarse con ventaja en problemas topológicos. Las piezas se definen utilizando una ecuación diferencial que apareció a finales del siglo XVIII, en

especial en el trabajo de Pierre-Simon de Laplace. En consecuencia, se le llama ecuación de Laplace. La investigación más importante de Laplace era en mecánica celeste, el movimiento y la forma de los planetas, lunas, cometas y estrellas. En 1783 estaba trabajando sobre la forma detallada de la Tierra. Para entonces se sabía que la Tierra no es una esfera perfecta sino que está achatada por los polos para formar un esferoide oblato, como un balón de playa en el que alguien está sentado. Pero incluso esa descripción deja escapar algo del detalle fino. Laplace encontró un método para calcular la forma con cualquier precisión requerida basado en una magnitud física que representa el campo gravitatorio de la Tierra: no el campo propiamente dicho sino su potencial gravitatorio. Este es una medida de la energía contenida en la gravitación, una magnitud numérica definida en cada punto del espacio. La fuerza de la gravedad actúa en la dirección que hace que el potencial disminuya al ritmo más rápido, y la magnitud de la fuerza es el ritmo de disminución.

El potencial satisface la ecuación de Laplace: a grandes rasgos, esto dice que en ausencia de materia —es decir, en el vacío— el valor medio del potencial sobre una esfera muy pequeña es igual a su valor en el centro de la esfera. Es una especie de democracia: el valor de uno es el promedio de los valores de sus vecinos. Cualquier solución de la ecuación de Laplace se llama una función armónica. Tipos especiales de Hodge de clases de cohomología son aquellos que guardan una relación particular con las funciones armónicas. La teoría de Hodge, el estudio de estos tipos, abrió una profunda y

maravillosa área de las matemáticas: relaciones entre la topología de un espacio y una ecuación diferencial especial en dicho espacio.

Así que ahora lo tenemos. La conjetura de Hodge postula una conexión profunda y poderosa entre tres de los pilares de las modernas matemáticas: álgebra, topología y análisis. Tomar cualquier variedad. Para entender su forma (topología, que lleva a clases de Hodge por vía de ecuaciones diferenciales) escoger ejemplos especiales de estas (análisis, que lleva a clases de Hodge por vía de ecuaciones diferenciales). Estos tipos especiales de clases de cohomología pueden realizarse utilizando subvariedades (álgebra: introducir algunas ecuaciones extra y examinar ciclos algebraicos). Es decir, para resolver el problema de topología «¿qué forma tiene esto?» para una variedad, se convierte la pregunta en análisis y luego se resuelve utilizando álgebra.

¿Por qué eso es importante? La conjetura de Hodge es una propuesta de añadir dos nuevas herramientas a la caja de herramientas de la geometría algebraica: invariantes topológicos y ecuación de Laplace. No es en realidad una conjetura sobre un teorema matemático: es una conjetura sobre nuevos tipos de herramientas. Si la conjetura es cierta, esas herramientas adquieren al momento nueva importancia y potencialmente pueden ser utilizadas para responder a una cadena sin fin de preguntas. Por supuesto, podría resultar que sea falsa. Eso sería desagradable, pero es mejor entender las limitaciones de una herramienta que seguir dándose en los dedos con ella.

Ahora que apreciamos la naturaleza de la conjetura de Hodge, podemos examinar las pruebas a su favor. ¿Qué sabemos? Bien poco.

En 1924, antes de que Hodge hiciera su conjetura, Solomon Lefschetz demostró un teorema que se reduce a la conjetura de Hodge para la cohomología en dimensión 2 de cualquier variedad. Con algo de topología algebraica rutinaria esto implica la conjetura de Hodge para variedades de dimensión 1, 2 y 3. Para variedades de dimensiones más altas solo se conocen unos pocos casos especiales de la conjetura de Hodge.

Hodge enunció originalmente su conjetura en términos de etiquetas enteras. En 1961 Michael Atiyah y Friedrich Hirzebruch demostraron que en dimensiones más altas esta versión de su conjetura es falsa. Por eso hoy interpretamos la conjetura de Hodge utilizando etiquetas racionales. A favor de esta versión hay alguna evidencia alentadora. La prueba más fuerte a su favor es que, *sin* dar por hecha la conjetura, se ha demostrado una de sus consecuencias más profundas, un teorema todavía más técnico conocido como «algebraicidad de lugares geométricos de Hodge». Eduardo Cattani, Pierre Deligne y Aroldo Kaplan encontraron dicha demostración en 1995.

Para finalizar, hay una conjetura atractiva en teoría de números que es análoga a la conjetura de Hodge. Se denomina conjetura de Tate, por John Tate, y vincula la geometría algebraica con la teoría de Galois, el círculo de ideas que demostraban que no hay fórmula algebraica para resolver ecuaciones polinómicas de grado 5. Su

formulación es técnica, e incluye otra versión más de cohomología. Hay razones independientes para esperar que la conjetura de Tate pueda ser cierta, pero su estatus está abierto por el momento. Pero al menos hay un pariente razonable de la conjetura de Hodge, incluso si de momento parece igualmente intratable.

La conjetura de Hodge es una de esas enojosas afirmaciones matemáticas para las que las pruebas a favor o en contra no son muy amplias ni especialmente convincentes. Sin duda, hay peligro de que la conjetura pueda ser falsa. Quizá haya una variedad con un millón de dimensiones que refute la conjetura de Hodge, por razones que se reducen a series de cálculos desestructurados, tan complicados que nadie podría realizarlos jamás. Si es así, la conjetura de Hodge podría ser falsa por razones que en realidad carecen de interés —sencillamente resulta no ser cierta— pero que en la práctica resultan imposibles de refutar. Conozco a algunos geómetras algebraicos que lo sospechan. Si es así, ese millón de dólares estará a salvo en un futuro previsible.

Capítulo 16
¿Qué viene ahora?

Hacer predicciones es muy difícil, especialmente sobre el futuro[87], como se supone que han dicho el físico Niels Bohr, ganador del premio Nobel, y el jugador de béisbol y director de equipo Yogi Berra[88]. Tenga en cuenta que Berra también dijo: «Nunca dije la mayoría de las cosas que dije». Supuestamente. Arthur C. Clarke, famoso por su ciencia ficción y por la película *2001: Una odisea del espacio* y sus secuelas, también era un futurólogo: escribió libros que predecían el futuro de la tecnología y de la sociedad. Entre las muchas predicciones en su *Perfiles del futuro* de 1962, figuran:

Comprensión de los lenguajes de ballenas y delfines para 1970.

Generación de energía por fusión para 1990.

Detección de ondas de gravedad para 1990.

Colonización de planetas para 2000.

Ninguna de estas ha sucedido todavía. Por otra parte, él sí tuvo algunos aciertos:

[87] Bohr puede haber puesto el dedo en la llaga. Las teorías científicas se ponen a prueba por sus predicciones, pero pocas de estas predicen el futuro. La mayoría son enunciados si/entonces: si se hace pasar luz a través de un prisma, se descompondrá en colores. La «predicción» no dice cuándo sucederá esto. Por ello, paradójicamente, podemos hacer predicciones sobre el tiempo sin predecir el tiempo. «Si el aire caliente procedente de un ciclón se junta con aire frío, entonces nevará» es una predicción científica, pero no un pronóstico.

[88] La cita o una variante parecida ha sido atribuida a unas treinta fuentes diferentes, incluidos Sam Goldwyn, Woody Allen, Winston Churchill y Confucio. Véase http://www.larry.denenberg.com/predictions.html.

Aterrizajes en planetas para 1980 (aunque quizá él se refiriera a aterrizajes de seres humanos).
Máquinas de traducir para 1970 (un poco prematuro, pero ahora existen en Google).
Radio personal para 1990 (los teléfonos móviles funcionan así).

También predijo que habría una biblioteca mundial para 2000, y esto puede estar más cerca de lo que pensábamos hace unos años, porque esta es una de las muchas funciones de internet. Con la llegada de la computación en la nube, podemos terminar utilizando todos el mismo ordenador gigante.

Él erró sobre algunas de las tendencias más importantes, tales como la aparición del ordenador y la ingeniería genética, aunque lo predijo para 2030. Con este desigual registro de Clarke como advertencia, sería temerario predecir en detalle el futuro de los grandes problemas matemáticos. Sin embargo, puedo hacer algunas conjeturas fundamentadas, con la casi certeza de que la mayoría de ellas resultarán erróneas.

En la introducción mencioné la lista de Hilbert en 1900 de 23 grandes problemas. La mayoría están ahora resueltos, y su valiente grito de guerra «Debemos saber, sabremos» puede parecer reivindicado. Sin embargo, él dijo también que «en matemáticas no hay *ignorabimus* [ignoraremos]» y Kurt Gödel asestó el golpe de gracia a esta idea con su teorema de incompletitud: algunos problemas matemáticos pueden no tener solución dentro del marco

lógico habitual de las matemáticas. No es solo que sean imposibles, como cuadrar el círculo: pueden ser indecidibles, lo que significa que no existe demostración ni existe refutación. Este podría ser el destino de algunos de los grandes problemas actualmente no resueltos. Yo me sorprendería si la hipótesis de Riemann fuera uno de estos, y me asombraría si alguien pudiera demostrar que es indecidible incluso si lo fuera. Por el contrario, el problema P/NP muy bien podría resultar indecidible, o satisfacer alguna otra variación técnica sobre el tema de «no puede hacerse». Tiene ese tipo de… bueno, *aroma*.

Yo sospecho que para finales del siglo XXI tendremos demostraciones de la hipótesis de Riemann, la conjetura de Birch-Swinnerton-Dyer y la hipótesis del hueco de masa, junto con refutaciones de la conjetura de Hodge y de la regularidad de las soluciones de la ecuación de Navier-Stokes en tres dimensiones. Y espero que P/NP siga sin estar resuelto en 2100, pero sucumba en algún momento del siglo XXII. Así que seguramente alguien refutará mañana la hipótesis de Riemann y demostrará que P es diferente de NP la próxima semana.

Estoy en terreno más seguro con observaciones generales, porque podemos aprender de la historia. Por eso estoy razonablemente confiado en que para cuando se hayan resuelto los siete problemas del milenio, muchos de ellos se verán como curiosidades históricas menores. «Oh, ellos solían pensar que eran importantes, ¿no es verdad?». Esto es lo que sucedió con alguno de los problemas de la lista de Hilbert. También puedo estar seguro de que en menos de

cincuenta años habrán nacido varias áreas importantes de las matemáticas que no existen hoy. Entonces se verá que unos pocos ejemplos básicos y algunos teoremas rudimentarios en dichas áreas existían mucho antes, pero nadie se dio cuenta de que estos fragmentos aislados eran claves de nuevas áreas profundas e importantes. Esto es lo que sucedió con la teoría de grupos, el álgebra matricial, los fractales y el caos. No dudo de que volverá a suceder, porque es una de las maneras estándar en que se desarrollan las matemáticas.

Estas nuevas áreas surgirán gracias a dos factores principales. Emergerán de la estructura interna de las propias matemáticas o serán respuestas a nuevas preguntas sobre el mundo exterior —a menudo ambos combinados—. Como el proceso en tres pasos de Poincaré para la solución de problemas —preparación, incubación e iluminación—, la relación entre las matemáticas y sus aplicaciones no es una única transición: la ciencia plantea un problema, las matemáticas lo resuelven, hecho. En su lugar, encontramos una intrincada red de intercambios de preguntas e ideas a medida que nuevas matemáticas desencadenan más experimentos u observaciones o teorías, que a su vez motivan nuevas matemáticas. Y cada nodo de esta red resulta ser, en un examen más cercano, una red más pequeña del mismo tipo.

Hay más mundo exterior que lo que solía haber. Hasta tiempos recientes la principal fuente externa de inspiración para las matemáticas eran las ciencias físicas. Algunas otras áreas tuvieron su papel: biología y sociología influyeron en el desarrollo de la

probabilidad y la estadística, y la filosofía tuvo un gran efecto en la lógica matemática. En el futuro veremos crecer las contribuciones de la biología, la medicina, la computación, las finanzas, la economía, la sociología y muy posiblemente la política, la industria cinematográfica y el deporte. Sospecho que algunos de los primeros nuevos grandes problemas vendrán de la biología, porque ese vínculo está ahora firmemente establecido. Una tendencia es una explosión en nuestra capacidad de reunir datos biológicos y bioquímicos; genomas pequeños pueden ahora secuenciarse utilizando un aparato del tamaño de un lápiz de memoria, basado en tecnología de nanoporos, por ejemplo. Rápidamente seguirán los grandes genomas utilizando esta tecnología u otra diferente, mucha de la cual ya existe.

Estos desarrollos son factores de cambio potenciales, pero necesitamos tener métodos mejores para entender lo que implican los datos. La biología no trata en realidad de los datos como tales. Trata de procesos. La evolución es un proceso, y también lo son la división celular, el crecimiento de un embrión, la aparición del cáncer, el movimiento de una multitud, el funcionamiento del cerebro y la dinámica del ecosistema global. La mejor manera que conocemos hoy día de tomar los ingredientes básicos de un proceso y deducir cómo actúa es las matemáticas. Por eso habrá grandes problemas de nuevos tipos: cómo se manifiesta la dinámica en presencia de información organizativa compleja pero específica (secuencias de ADN); cómo los cambios genéticos conspiran con el ambiente para guiar la evolución; cómo las reglas para el

crecimiento, división, movilidad, adhesividad y muerte de las células dan su forma a los organismos en desarrollo; cómo el flujo de electrones y sustancias químicas en una red de células nerviosas determina lo que puede percibir o cómo actuará.

La computación es otra fuente de nuevas matemáticas que ya tiene un historial. Normalmente se considera como una herramienta para hacer matemáticas, pero las matemáticas son asimismo una herramienta para entender y estructurar las computaciones. Este intercambio en dos direcciones se está haciendo cada vez más importante para la salud y desarrollo de ambas áreas, que incluso pueden fusionarse en algún momento en el futuro. Algunos matemáticos piensan que nunca se debería haber permitido que se separaran. Entre las muchas tendencias visibles aquí, la cuestión de enormes conjuntos de datos viene de nuevo a la mente. No solo se relaciona con el ejemplo del ADN antes mencionado sino también con la predicción de terremotos, la evolución, el clima global, el mercado de valores, las finanzas internacionales y las nuevas tecnologías. El problema está en utilizar grandes cantidades de datos para poner a prueba y refinar los modelos matemáticos del mundo real, de modo que nos dan un control genuino sobre sistemas muy complejos.

La predicción en la que tengo más confianza es en algunos aspectos negativa, pero es también una afirmación de la continua creatividad de la comunidad matemática. Todos los matemáticos que investigan piensan, en ocasiones, que su disciplina tiene una mente propia. Los problemas trabajan de la forma en que los quieren las

matemáticas, y no cómo los quieren los matemáticos. Podemos escoger qué preguntas plantear, pero no podemos escoger qué respuestas obtenemos. Esta sensación se relaciona con dos escuelas principales de pensamiento sobre la naturaleza de las matemáticas. Los platónicos piensan que las «formas ideales» de las matemáticas tienen algún tipo de existencia independiente «ahí fuera», en algún reino distinto del mundo físico. (Hay formas más sutiles de decirlo y que probablemente suenan más razonables, pero esa es la esencia). Otros ven las matemáticas como una construcción humana compartida. Pero a diferencia de la mayoría de tales cosas —el sistema legal, el dinero, la ética, la moralidad— las matemáticas son una construcción con un fuerte esqueleto lógico. Hay varias restricciones severas sobre qué afirmaciones pueden o no compartirse con cualquier otro. Son estas restricciones las que dan la impresión de que las matemáticas tienen su propia agenda y crean la sensación en la mente de los matemáticos de que las propias matemáticas *existen* fuera del dominio de la actividad humana. El platonismo, pienso yo, no es una descripción de lo que son las matemáticas. Es una descripción de lo que las matemáticas parecen cuando uno las está haciendo. Es como la vívida sensación de «rojo» que experimentamos cuando vemos una rosa, sangre o un semáforo. Los filósofos llaman *qualia* (singular: *quale*) a estas sensaciones, y algunos piensan que nuestra sensación de libre albedrío es en realidad un quale de la manera del cerebro de tomar decisiones. Cuando decidimos entre alternativas tenemos la sensación de que hacemos una elección genuina, sea o no

realmente determinista en algún sentido la dinámica del cerebro. De modo análogo, el platonismo es un *quale* de tomar parte en una construcción humana compartida dentro de un rígido marco de deducción lógica.

Por ello puede parecer que las matemáticas tienen una mente propia, incluso si son creadas por un conjunto de mentes humanas. La historia nos dice que la mente matemática es, en este sentido, más innovadora y sorprendente que lo que cualquier mente humana individual pueda predecir. Todo lo cual es una manera complicada de llegar a mi punto principal: una cosa que podemos predecir con seguridad sobre el futuro de las matemáticas es que serán impredecibles. Las preguntas matemáticas más importantes del próximo siglo emergerán como consecuencias naturales, quizá incluso inevitables, de nuestra creciente comprensión de lo que actualmente creemos que son los grandes problemas de las matemáticas. Sin embargo, casi con seguridad serán preguntas que hoy día no podemos concebir. Esto es justo y adecuado, y deberíamos celebrarlo.

Capítulo 17
Doce para el futuro

No quiero dejarle con la impresión de que la mayoría de los problemas matemáticos han sido resueltos, aparte de los singularmente difíciles. Investigar en matemáticas es como explorar un nuevo continente. A medida que se expande el área que conocemos, la frontera que rodea lo desconocido se hace más grande. No estoy sugiriendo que cuantas más matemáticas descubrimos menos sabemos; estoy diciendo que cuantas más matemáticas descubrimos, más nos damos cuenta de lo que no sabemos. Pero lo que no sabemos cambia con el paso del tiempo: algunos viejos problemas desaparecen mientras se añaden otros nuevos. Por el contrario, lo que sabemos se hace más grande, salvo el ocasional documento perdido.

Para darle un minúsculo indicio de lo que actualmente *no* sabemos, además de los grandes problemas ya discutidos, he aquí doce problemas no resueltos que llevan desconcertando bastante tiempo a los matemáticos de todo el mundo. Los he escogido de modo que las preguntas sean fáciles de entender. Como ha sido ampliamente demostrado, eso no tiene consecuencias sobre lo fácil que pueda ser encontrar las respuestas. Algunos de estos problemas pueden resultar grandes: ello dependerá sobre todo de los métodos que se ingenien para resolverlos y de adónde lleven, y no de la respuesta como tal.

El problema de Brocard

Para cualquier número natural n, su factorial $n!$ es el producto

$$n\times(n - 1)\times(n - 2)\times\ldots\times3\times2\times1$$

Este es el número de formas diferentes de ordenar n objetos. Por ejemplo, el alfabeto inglés con 26 letras puede ordenarse de

$$26! = 403.291.461.126.605.635.584.000.000$$

maneras diferentes[v]. En artículos escritos de 1876 y 1888, Henri Brocard advirtió que

$$4! + 1 = 24 + 1 = 25 = 5^2$$
$$5! + 1 = 120 + 1 = 121 = 11^2$$
$$7! + 1 = 5040 + 1 = 5041 = 71^2$$

son todos cuadrados perfectos. No encontró otros factoriales que dieran cuadrados perfectos cuando se les sumaba 1, y preguntó si existía alguno. El genio indio autodidacta Srinivasa Ramanujan planteó independientemente la misma pregunta en 1913. Bruce Berndt y William Galway utilizaron un ordenador en 2000 para mostrar que no existen más soluciones para factoriales de números hasta 1000 millones.

Números perfectos impares

Un número es perfecto si es igual a la suma de todos sus divisores propios (es decir, números que lo dividen exactamente, excluido el propio número). Ejemplos son:

$$6 = 1 + 2 + 3$$
$$28 = 1 + 2 + 4 + 7 + 14$$

Euclides demostró que si 2n - 1 es primo, entonces $2^{n-1}(2n - 1)$ es perfecto. Los ejemplos anteriores corresponden a $n = 2, 3$. Los primos de esta forma se llaman primos de Mersenne, y se conocen 47 de ellos, de los que el mayor hasta la fecha es $2^{43.112.609} - 1$, que es también el mayor primo conocido[89]. Euler demostró que todos los números perfectos pares deben ser de esta forma, pero nadie ha encontrado nunca un número perfecto impar, ni demostrado que no pueden existir. Pomerance ha concebido un argumento no riguroso que indica que no existen. Cualquier número perfecto impar debe satisfacer varias condiciones restrictivas. Debe ser al menos 10^{300}, debe tener un factor primo mayor que 10^8, su segundo mayor factor primo debe ser al menos 10^4, y debe tener al menos 75 factores primos y al menos 12 factores primos distintos.

La conjetura de Collatz

Tomemos un número entero. Si es par, dividámoslo por 2. Si es impar, multipliquémoslo por 3 y sumemos 1. Repitámoslo indefinidamente. ¿Qué sucede?

[89] Para la información más reciente, véase «The Prime Pages»: http:// primes.utm.edu.

Por ejemplo, empecemos con 12. Los números sucesivos son:

$$12 \rightarrow 6 \rightarrow 3 \rightarrow 10 \rightarrow 5 \rightarrow 16 \rightarrow 8 \rightarrow 4 \rightarrow 2 \rightarrow 1$$

después de lo cual la secuencia

$$4 \rightarrow 2 \rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow 1$$

se repite indefinidamente. La conjetura de Collatz afirma que el mismo resultado final se obtiene cualquiera que sea el número del que partimos. El nombre se debe a Lothar Collatz quien dio con ella en 1937, pero tiene muchos otros nombres: conjetura $3n + 1$, problema de granizo, conjetura de Ulam, problema de Kakutani, conjetura de Thwaites, algoritmo de Hasse y problema de Syracuse.
Lo que hace el problema difícil es que a menudo los números pueden explotar. Por ejemplo, si partimos de 27 entonces la secuencia crece hasta 9232; incluso así, finalmente se reduce a 1 después de 111 pasos. Las simulaciones mediante ordenador verifican la conjetura para todos los números iniciales hasta $5{,}764 \times 10^{18}$. Se ha demostrado que no existen ciclos distintos de 4 → 2 → 1 que incluyan menos de 35.400 números. La posibilidad de que algún número inicial lleve a una secuencia que contenga números cada vez mayores, separados por números más pequeños, no ha sido descartada. Ilia Krasikov y Jeffrey Lagarias han demostrado que para valores iniciales hasta n, al menos una

constante multiplicada por $n^{0,84}$ de ellos llega finalmente a 1. Por lo tanto, las excepciones, si existen, son raras[90].

Existencia de cuboides perfectos

Esto toma como punto de partida la existencia de, y la fórmula para, tripletas pitagóricas, y traslada el problema a la tercera dimensión. Un ladrillo de Euler es un cuboide —un bloque con forma de ladrillo— con lados enteros, todas cuyas caras tienen diagonales enteras. El ladrillo de Euler más pequeño fue descubierto en 1719 por Paul Halcke. Sus aristas son 240, 117 y 4; las diagonales de las caras son 267, 244 y 125. Euler encontró fórmulas para tales ladrillos, análogas a la fórmula para las tripletas pitagóricas, pero estas no dan todas las soluciones.

No se sabe si existe un cuboide perfecto: es decir, si existe un ladrillo cuya diagonal principal, que atraviesa el interior del ladrillo desde una esquina a la opuesta, también tiene longitud entera. (Hay cuatro de estas diagonales pero todas tienen la misma longitud). Se sabe que las fórmulas de Euler no pueden proporcionar un ejemplo. Un ladrillo semejante, si existe, debe satisfacer varias condiciones; por ejemplo, al menos una arista debe ser un múltiplo de 5, otra debe ser un múltiplo de 7, otra debe ser un múltiplo de 11 y otra debe ser un múltiplo de 19. Las búsquedas por ordenador han mostrado que uno de los lados debe ser al menos un billón.

Hay algunas aproximaciones casi exactas. El ladrillo con lados 672, 153 y 104 tiene una diagonal principal entera y dos de las tres

[90] Ilia Krasikov y Jeffrey C. Lagarias, «Bounds for the 3x + 1 problemas using difference inequalities», *Acta Arithmetica*, n.º 109 (2003), pp. 237-258.

longitudes para las diagonales de las caras son también enteras. En 2004 Jorge Sawyer y Clifford Reiter demostraron que existen paralelepípedos perfectos[91]. Un paralelepípedo es como un cuboide pero sus caras son paralelogramos. Por eso está inclinado. Las aristas tienen longitudes 271, 106 y 103; las diagonales de la cara menor tienen longitudes 101, 266 y 255; las diagonales de la cara mayor tienen longitudes 183, 312 y 323; y las diagonales del cuerpo tienen longitudes 374, 300, 278 y 272.

La conjetura del corredor solitario

Esta procede de una abstrusa área de las matemáticas conocida como teoría de aproximación diofántica, y fue formulada por Jörg Wills en 1967. Luis Goddyn acuñó el nombre en 1998. Supongamos que n corredores recorren un trayecto circular de longitud unidad con velocidad uniforme. ¿Estará cada corredor solitario —es decir, estará a más de una distancia $1/n$ de todos los demás corredores— en algún instante? Diferentes instantes para diferentes corredores, por supuesto. La conjetura es que la respuesta es siempre «sí», y se ha demostrado cuando $n = 4, 5, 6$ y 7.

[91] Jorge F. Sawyer y Clifford A. Reiter, «Perfect parallelepipeds exist». arXiv:0907.0220 (2009).

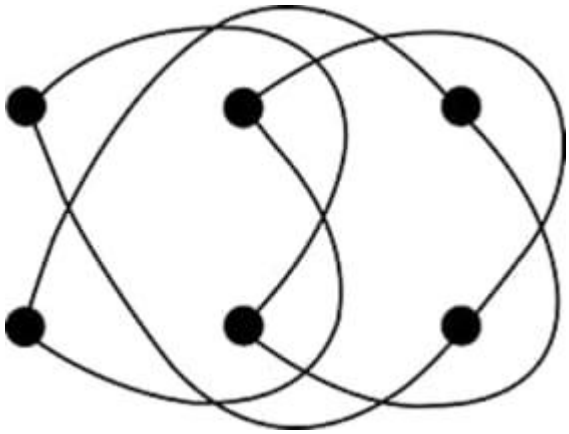


Figura 48. Ejemplo de un thrackle.

La conjetura del *thrackle* de Conway

Un *thrackle* es una red trazada en el plano de modo que cada dos aristas se encuentran exactamente una vez (véase Figura 48). Pueden encontrarse en un punto común (nodo, vértice) o pueden cruzarse en puntos interiores, pero no ambas cosas. Si se cruzan, deben hacerlo transversalmente; es decir, tampoco pueden permanecer una siempre en el mismo lado de la otra (lo que podría suceder si, por ejemplo, son tangentes). En un trabajo no publicado, John Horton Conway conjeturó que en cualquier *thrackle* el número de líneas es menor o igual que el número de puntos. En 2011 Radoslav Fulek y János Pach demostraron que todo *thrackle* con n puntos tiene como máximo $1{,}428n$ líneas[92].

Irracionalidad de la constante de Euler

No se conoce ninguna fórmula en «forma cerrada» para la suma de la serie armónica

[92] R. Fulek y J. Pach, «A computational approach to Conway's thrackle conjecture», *Computational Geometry*, n.º 44 (2011), pp. 345-355.

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \ldots + \frac{1}{n}$$

y probablemente no existe tal fórmula. Sin embargo, hay una aproximación excelente: cuando *n* aumenta, *H*n se aproxima cada vez más a log *n* + γ. Aquí γ es la constante de Euler, con un valor numérico de aproximadamente 0,5772156649. Euler estableció esta fórmula en 1734, y Lorenzo Mascheroni estudió la constante en 1790. Ninguno de ellos utilizó el símbolo γ.

La constante de Euler es uno de esos números extraños que surgen a veces en matemáticas, como π y e, que salen por todas partes, pero parecen ser criaturas autónomas, no expresables de ninguna manera clara en términos de números más sencillos. Vimos en el capítulo 3 que tanto π como e son trascendentes: no son soluciones de ninguna ecuación algebraica con coeficientes enteros. En particular, son irracionales: no son fracciones exactas. Es generalmente admitido que la constante de Euler es trascendente, pero ni siquiera sabemos con certeza que sea irracional. Si $\gamma = p/q$ para enteros *p* y *q*, entonces *q* es al menos $10^{242.080}$.

La constante de Euler es importante en muchas áreas de las matemáticas, que van desde la función zeta de Riemann a la teoría cuántica de campos. Aparece en muchos contextos y se manifiesta en muchas fórmulas. Es escandaloso que no podamos decidir si es racional.

Campos de números cuadráticos reales

En el capítulo 7 vimos que algunos campos de números algebraicos tienen factorización en primos única y otros no la tienen. Los campos de números algebraicos mejor entendidos son los cuadráticos, obtenidos tomando la raíz cuadrada de un número d que no es un cuadrado perfecto; de hecho, no tiene factores primos cuadrados. El correspondiente anillo de enteros algebraicos consiste entonces en todos los números de la forma $a + b\sqrt{d}$, donde a y b son enteros si d no es de la forma $4k + 1$, y son o bien enteros o son ambos enteros impares divididos por 2, si d es de esa forma.

Cuando d es negativo, es sabido que la factorización en primos es única para exactamente nueve valores:

-1, -2, -3, -7, -11, -19, -43, -67 y -163.

Demostrar la unicidad en estos casos es relativamente sencillo, pero encontrar si hay otros es mucho más difícil. En 1934 Hans Heilbronn y Edward Linfoot mostraron que a lo sumo puede añadirse a la lista un entero negativo más. Kurt Heegner dio una demostración de que la lista es completa en 1952, pero se pensaba que había una laguna. En 1967 Harold Stark dio una demostración completa, observando que no difería de forma significativa de la de Heegner, es decir, la laguna no era importante. Prácticamente al mismo tiempo, Alan Baker encontró una demostración diferente.

El caso cuando d es positivo es muy diferente. La factorización es única para muchos más valores de d. Hasta 50, dichos valores son 2, 3, 5, 6, 7, 11, 13, 14, 17, 19, 21, 22, 23, 29, 31, 33, 37, 38, 41,

43, 46 y 47, y cálculos por ordenador revelan muchos más. Por todo lo que sabemos, puede haber infinitos d positivos para los que el correspondiente campo de números cuadráticos tiene factorización única. Un análisis heurístico de Cohen y Lenstra sugiere que aproximadamente tres cuartos de todos los d positivos definen campos de números con factorización única. Los resultados obtenidos por ordenador coinciden con esta estimación. El problema está en demostrar que estas observaciones son correctas.

La hormiga de Langton

A medida que transcurre el siglo XXI se ha hecho cada vez más evidente que algunas de las técnicas tradicionales de la modelización matemática son incapaces de tratar las complejidades de los problemas a que se enfrenta la humanidad, tales como el sistema financiero mundial, la dinámica de los ecosistemas y el papel de los genes en el crecimiento de los organismos vivos. Muchos de estos sistemas incluyen grandes números de agentes —personas, compañías, organismos, genes— que interaccionan entre sí. Estas interacciones pueden modelarse a menudo con mucha exactitud utilizando reglas simples. Durante los últimos treinta años ha aparecido un nuevo tipo de modelo que trata de abordar frontalmente el comportamiento de sistemas con muchos agentes. Para entender cómo se moverán cien mil personas en un estadio, por ejemplo, no se las promedia para crear una especie de fluido humano y preguntar cómo fluye. En su lugar, se construye un modelo para ordenador con cien mil agentes individuales, se

imponen reglas apropiadas y se realiza una simulación para ver qué hace esta multitud de ordenador. Este tipo de modelo se denomina sistema complejo.

Para darle una idea de esta nueva y fascinante área de las matemáticas voy a describir uno de los más simples sistemas complejos y explicar por qué no lo entendemos plenamente. Se denomina la hormiga de Langton. Christopher Langton fue uno de los miembros iniciales del Instituto de Santa Fe, fundado en 1984 por los científicos George Cowan, Murray Gell-Mann y otros para promover la teoría y aplicaciones de los sistemas complejos. Langton inventó su hormiga en 1986. Técnicamente es un autómata celular, un sistema de celdas en una malla cuadrada cuyos estados se muestran mediante colores. En cada paso de tiempo el color de una celda cambia de una forma que depende de los colores de sus vecinas.

Las reglas son absurdamente simples. La hormiga vive en una malla cuadrada infinita de celdas, que al principio son todas blancas. Lleva un inagotable bote de pintura negra de secado rápido y otro inagotable bote de pintura blanca de secado rápido. Puede poner cara al norte, el este, el sur o el oeste; por simetría podemos suponer que empieza de cara al norte. En cada instante mira el color del cuadrado que ocupa y lo cambia de blanco a negro, o de negro a blanco, utilizando sus botes de pintura. Si su cuadrado era blanco, entonces gira 90 ° a la derecha y da un paso adelante. Si su cuadrado era negro, entonces gira 90 ° a la izquierda y da un paso adelante. Luego repite este comportamiento indefinidamente.

Si simulamos la hormiga[93], empieza pintando dibujos sencillos y bastante simétricos de cuadrados negros y blancos. De cuando en cuando vuelve a un cuadrado que ya ha visitado, pero su recorrido no se cierra en un lazo porque el color en dicho cuadrado ha cambiado, de modo que cuando repite la visita gira en el otro sentido. Conforme continúa la simulación, el dibujo de la hormiga se hace caótico y aleatorio. No hay una pauta discernible: básicamente es solo una mezcolanza. En esa etapa es razonable imaginar que este comportamiento caótico continúa de modo indefinido. Después de todo, cuando la hormiga revisita una región caótica hará una serie caótica de giros y repinturas. Si siguiéramos adelante con la simulación, los siguientes diez mil pasos parecerían justificar esa conclusión. Sin embargo, si se continúa aparece una pauta. La hormiga entra en un ciclo repetitivo de 104 pasos, al cabo de los cuales se ha movido tres cuadrados en diagonal. Entonces pinta una ancha banda diagonal de celdas negras y blancas, llamada una autopista, que continúa indefinidamente (véase Figura 49).

[93] http://en.wikipedia.org/wiki/Langton9627_ant.

Figura 49. Autopista de la hormiga de Langton.

Todo lo descrito hasta ahora puede demostrarse con todo rigor, simplemente dando la lista de pasos que da la hormiga. La demostración sería muy larga —una lista de cien mil pasos—, aunque seguiría siendo una demostración. Pero las matemáticas se hacen más interesantes si planteamos una pregunta algo más general. Supongamos que antes de que empiece la hormiga, pintamos de negro un número finito de cuadrados. Podemos escoger estos cuadrados de la manera que queramos: puntos al azar, un rectángulo sólido, la Mona Lisa. Podemos utilizar un millón de ellos, o mil millones, pero no infinitos. ¿Qué sucede?

Las excursiones iniciales de la hormiga cambian de forma espectacular cada vez que encuentra uno de nuestros nuevos cuadrados negros. Puede moverse por todas partes, trazando formas intrincadas o redibujándolas... Pero en todas las simulaciones realizadas, independientemente de cuál pudiera ser la configuración

inicial, la hormiga al fin se dispone a construir su autopista, utilizando el mismo ciclo de 104 pasos. ¿Siempre sucede esto? ¿Es la autopista el único «atractor» para la dinámica de la hormiga? Nadie lo sabe. Es uno de los problemas básicos no resueltos de la teoría de la complejidad. Lo más que sabemos es que cualquiera que pueda ser la configuración inicial de celdas negras, la hormiga no puede permanecer para siempre dentro de una región acotada de la malla.

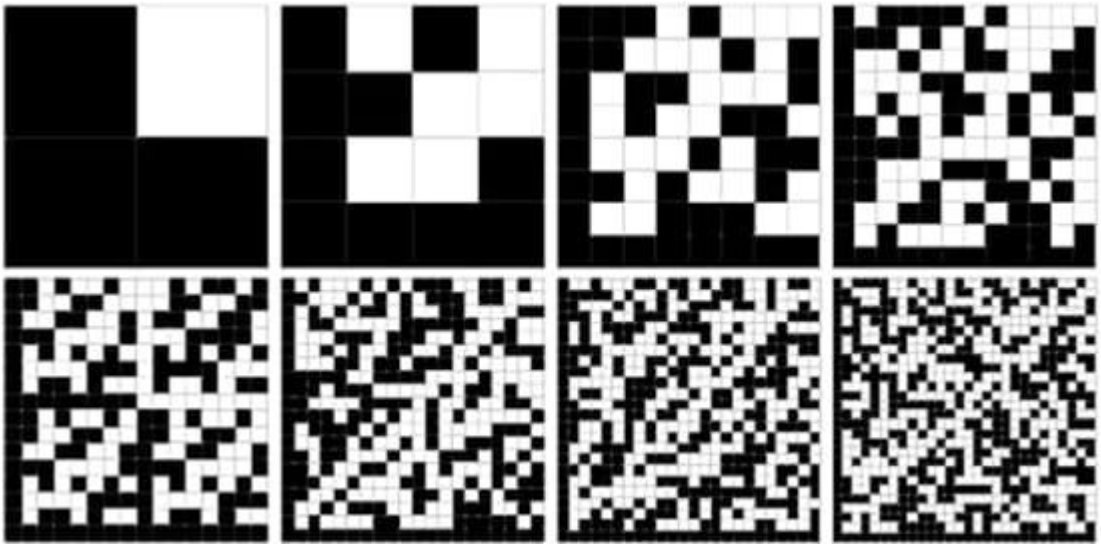


Figura 50. Matrices de Hadamard de tamaño 2, 4, 8, 12, 16, 20, 24 y 28. © Wolfram MathWorld.
http://mathworld.wolfram.com/HadamardMatrix.html.

La conjetura de la matriz de Hadamard

Una matriz de Hadamard, así llamada por Jacques Hadamard, es una formación cuadrada de ceros y unos tal que cualesquiera dos filas o columnas distintas coinciden en la mitad de sus entradas y se diferencian en la otra mitad. Utilizando negro y blanco para

indicar 1 y 0, la Figura 50 muestra matrices de Hadamard de tamaño 2, 4, 8, 12, 16, 20, 24 y 28. Estas matrices aparecen en muchos problemas matemáticos, y en ciencia de la computación, especialmente en la teoría de la codificación. (En algunas aplicaciones, entre ellas la motivación original de Hadamard, los cuadrados blancos corresponden a -1, no a 0.)

Hadamard demostró que tales matrices existen solo cuando $n = 2$ o n es un múltiplo de 4. El teorema de Paley de 1933 demuestra que una matriz de Hadamard existe siempre si el tamaño es un múltiplo de 4 e igual a $2^a(p^b + 1)$ donde p es un primo impar. Múltiplos de 4 no cubiertos por este teorema son 92, 116, 156, 172, 184, 188, 232, 236, 260, 268 y otros valores más grandes. La conjetura afirma que una matriz de Hadamard existe cuando quiera que el tamaño es un múltiplo de 4. En 1985 K. Sawade encontró una de tamaño 268; los otros números no cubiertos por el teorema de Paley ya habían sido tratados. En 2004 Hadi Kharaghani y Behruz Tayfeh-Rezaie encontraron una matriz de Hadamard de tamaño 428, y el tamaño más pequeño para el que no se conoce la respuesta es ahora 668.

La ecuación de Fermat-Catalan

Esta es la ecuación diofántica $x^a + y^b = z^c$ donde los exponentes a, b y c son enteros positivos. La llamaré ecuación de Fermat-Catalan porque sus soluciones están relacionadas con el último teorema de Fermat (véase el capítulo 7), y con la conjetura de Catalan (véase capítulo 6). Si a, b y c son pequeños, las soluciones enteras no nulas no son especialmente sorprendentes. Por ejemplo, si todos

son 2, entonces tenemos la ecuación pitagórica, que desde el tiempo de Euclides se sabe que tiene infinitas soluciones. De modo que el mayor interés está en los casos en que estos exponentes son grandes. La definición técnica de «grande» es que $s = 1/a + 1/b + 1/c$ es menor que 1. Solo se conocen diez soluciones grandes de la ecuación de Fermat-Catalan:

$$1 + 2^3 = 3^2 \qquad 17^7 + 76.271^3 = 21.063.928^2$$
$$2^5 + 7^2 = 3^4 \qquad 1.414^3 + 2.213.459^2 = 65^7$$
$$7^3 + 13^2 = 2^9 \qquad 9.262^3 + 15.312.283^2 = 113^7$$
$$2^7 + 17^3 = 71^2 \qquad 43^8 + 9.622^3 = 30.042.907^2$$
$$3^5 + 11^4 = 122^2 \qquad 33^8 + 159.034^2 = 15.613^3$$

La primera de estas se considera grande porque $1 = 1a$ para cualquier a, y $a = 7$ satisface la definición. La conjetura de Fermat-Catalan afirma que la ecuación de Fermat-Catalan tiene solamente un número finito de soluciones enteras, sin un factor común, cuando s es grande. El resultado más importante fue demostrado en 1997 por Henri Darmon y Loïc Merel: no existen soluciones en las que $c = 3$ y a y b son iguales y mayores que 3. Poco más se sabe. Progresos adicionales dependen de una nueva y fascinante conjetura que viene a continuación.

La conjetura ABC

En 1983 Richard Mason advirtió que se había ignorado un caso del último teorema de Fermat: potencias primeras. Es decir, consideremos la ecuación $a + b = c$.

A primera vista esta idea es completamente absurda. No hay que saber mucha álgebra para resolver esta ecuación para cualquiera de las tres variables en términos de las otras dos. Por ejemplo $a = c - b$. Lo que cambia todo el juego, sin embargo, es el contexto. Mason se dio cuenta de que todo se hacía mucho más profundo si planteamos las preguntas correctas sobre a, b y c. El resultado de esta idea extraordinaria fue una nueva conjetura en teoría de números con consecuencias de largo alcance. Podría resolver muchos problemas actualmente no resueltos y llevar a demostraciones mejores y más sencillas de algunos de los más grandes teoremas en teoría de números. Esta es la conjetura ABC, y tiene muchas pruebas numéricas a su favor. Se basa en una vaga analogía entre enteros y polinomios.

Euclides y Diofanto conocían una receta para tripletas pitagóricas, que ahora escribimos como una fórmula (véase capítulo 6). ¿Puede repetirse este truco con otras ecuaciones? En 1851 Joseph Liouville demostró que no existe tal fórmula para la ecuación de Fermat cuando la potencia es 3 o mayor. Mason aplicó un razonamiento similar a la ecuación más simple

$$a(x) + b(x) = c(x)$$

para tres polinomios. Es una idea escandalosa porque todas las soluciones pueden encontrarse utilizando álgebra elemental. El resultado principal, sin embargo, es elegante y nada obvio: si cada

polinomio tiene un factor que es un cuadrado, un cubo o una potencia superior, la ecuación no tiene soluciones.

Los teoremas sobre polinomios suelen tener análogos sobre enteros. En particular, polinomios irreducibles corresponden a números primos. El análogo natural en enteros del teorema de Mason sobre polinomios es como sigue. Supongamos $a + b = c$ donde a, b y c son enteros sin ningún factor común; entonces el número de factores primos de cada uno de los a, b y c es menor que el número de factores primos *distintos* de abc. Por desgracia, ejemplos sencillos muestran que esto es falso. En 1985 David Masser y Joseph Oesterlé modificaron el enunciado y propusieron una versión de esta conjetura que no estaba en conflicto con ningún ejemplo conocido. Su conjetura ABC muy bien puede ser la mayor pregunta abierta en teoría de números en el momento presente[94]. Si alguien demostrara mañana la conjetura ABC, muchos teoremas profundos y difíciles, comprobados en las últimas décadas con enorme intuición y esfuerzo, tendrían demostraciones nuevas y sencillas. Otra consecuencia sería la conjetura de Marshall Hall: la diferencia entre cualquier cubo perfecto y cualquier cuadrado perfecto tiene que ser bastante grande. Otra potencial aplicación de la conjetura ABC es al problema de Brocard, el primero de este capítulo. En 1993 Marius Overholt demostró que si la conjetura ABC es cierta, hay solo un número finito de soluciones a la ecuación de Brocard.

[94] La conjetura abc afirma: Para cualquier $\varepsilon > 0$ existe una constante $k\varepsilon > 0$ tal que si a, b y c son enteros positivos que no tienen ningún factor común mayor que 1, y $a + b = c$, entonces $c \leq k_\varepsilon P^{1+\varepsilon}$, donde P es el producto de todos los primos distintos que dividen abc.

Una de las consecuencias más interesantes de la conjetura ABC está relacionada con la conjetura de Mordell. Faltings la ha demostrado utilizando métodos sofisticados, pero su resultado sería aún más poderoso si conociéramos alguna información extra: una cota sobre el tamaño de las soluciones. Entonces existiría un algoritmo para encontrarlas todas. En 1991 Noam Elkies mostró que una versión específica de la conjetura ABC, en la que varias constantes que aparecen están acotadas, implica esta mejora sobre el teorema de Faltings. Laurent Moret-Bailly demostró que la recíproca es cierta, en un sentido muy fuerte. Cotas suficientemente fuertes sobre el tamaño de las soluciones de *tan solo* una ecuación diofántica, $y^2 = x^5 - x$, implican la conjetura ABC plena. Aunque no sea tan bien conocida como muchas otras conjeturas no resueltas, la conjetura ABC es indudablemente uno de los grandes problemas de las matemáticas. Según Grainville y Thomas Tucker, disponer de ella tendría «un extraordinario impacto sobre nuestra comprensión de la teoría de números. Demostrarla o refutarla sería asombroso[95]».

[95] Andrew Granville y Thomas J. Tucker, «It's as easy as abc», *Notices of the American Mathematical Society*, n.º 49 (2002), pp. 1224-1231. En septiembre de 2012 Shinichi Mochizuki anunció que había demostrado la conjetura ABC utilizando una aproximación radicalmente nueva a los fundamentos de la geometría algebraica. Los expertos están ahora comprobando su demostración de quinientas páginas, pero esto puede llevar mucho tiempo.

Glosario

Algoritmo. Un procedimiento especifico para resolver un problema, con garantía de que se detiene con una respuesta.

Análisis complejo. Análisis —cálculo infinitesimal con rigor lógico— realizado con funciones de valor complejo de una variable compleja.

Aritmética modular. Un sistema de aritmética en el que todos los múltiplos de cierto número específico, llamado el *módulo*, se tratan como si fueran cero.

Asintótico. Dos cantidades definidas en términos de una variable son asintóticas si su razón se hace cada vez más próxima a 1 a medida que la variable se hace arbitrariamente grande.

Autovalor (valor propio). Uno de un conjunto de números especiales asociado a un operador. Si el operador aplicado a un vector da un múltiplo de dicho vector, el múltiplo en cuestión es un autovalor.

Bola. Una esfera maciza; es decir, una esfera y su interior.

Bosón de Higgs. Una partícula fundamental cuya existencia explica por qué todas las partículas tienen masa. Su descubrimiento en el Gran Colisionador de Hadrones se anunció en julio de 2012.

Campo de velocidades. Una función que especifica una velocidad en cada punto del espacio. Por ejemplo, cuando fluye un fluido, su velocidad puede especificarse en cada punto, y normalmente difiere en puntos diferentes.

Campo electromagnético. Una función que especifica las intensidades y direcciones de los campos eléctrico y magnético en cualquier punto del espacio.

Caos. Comportamiento aparentemente aleatorio en un sistema determinista.

Característica de Euler. $C - A - V$, donde C es el número de caras en una triangulación de un espacio, A es el número de aristas y V es el número de vértices. Para un toro con g agujeros es igual a $2 - 2g$, cualquiera que pueda ser la triangulación.

Cero (de una función). Si f es una función, entonces x es un cero de f si $f(x) = 0$.

Ciclo. En topología: una combinación formal de lazos en una triangulación con etiquetas numéricas asociadas. En geometría algebraica: una combinación formal de subvariedades con etiquetas numéricas asociadas.

Clase de Hodge. Una clase de cohomología de ciclos sobre una variedad algebraica con propiedades analíticas especiales.

Clase E. Un algoritmo cuyo tiempo de ejecución, para una entrada de tamaño n, va con la n -ésima potencia de una constante.

Clase no P. No de clase P.

Clase NP. Un problema para el que una solución propuesta puede comprobarse (pero no necesariamente encontrarse) mediante un algoritmo de clase P.

Clase P. Un algoritmo cuyo tiempo de ejecución va como una potencia fija del tamaño de la entrada.

Coeficiente. En un polinomio tal como $6x^3 - 5x^2 + 4x - 7$, los coeficientes son los números 6, -5, 4, -7 que multiplican a las diversas potencias de *x*.

Configuración inevitable. Un miembro de una lista de redes, de los que al menos uno debe ocurrir en cualquier red en el plano.

Configuración reducible. Una parte de una red con la siguiente propiedad: si la red obtenida al eliminarla puede ser coloreada con cuatro colores, también puede serlo la red original.

Conjunto. Una colección de objetos (matemáticos). Por ejemplo, el conjunto de todos los números naturales.

Construcción por regla y compás. Cualquier construcción geométrica que puede realizarse utilizando una regla sin marcas y un compás (estrictamente: un par de compases).

Contraejemplo. Un ejemplo que refuta una afirmación. Así, 9 es un contraejemplo de la afirmación «todos los números impares son primos».

Constante de Euler. Un número especial denotado por γ, aproximadamente igual a 0,57721. (Véase nota 67).

Coordenada. Un número en una lista que determina la posición de un punto en un plano o en el espacio.

Coseno. Una función trigonométrica de un ángulo, definida por $\cos A = a/c$ en la Figura 51.

Figura 51. El coseno (a/c), seno (b/c) y tangente (a/b) de un ángulo *A*.

Cota superior. Un número específico del que se sabe con certeza que es mayor que cierta cantidad cuyo tamaño se está buscando.

Criminal mínimo. Un objeto matemático que no posee una propiedad deseada, y en cierto sentido es el más pequeño posible de tales objetos. Por ejemplo, un mapa que no puede colorearse con cuatro colores y que también tiene el menor número de regiones para las que esto puede ocurrir. Los criminales mínimos suelen ser hipotéticos, y el objetivo es demostrar que no pueden existir.

Cuadrado. Un número multiplicado por sí mismo. Por ejemplo, el cuadrado de 7 es 7×7 = 49, simbolizado por 7^2.

Cubo. Un número multiplicado por sí mismo y luego vuelto a multiplicar por sí mismo. Por ejemplo, el cubo de 7 es 7×7×7 =343. Normalmente se escribe como 7^3.

Curva elíptica. Una curva en el plano cuya ecuación tiene la forma $y^2 = ax^3 + bx^2 + cx + d$ en donde normalmente se supone que las constantes a, b, c, d, son racionales. (Véase Figura 27).

Curvatura. Una medida de cuánto se curva el espacio cerca de un punto dado. Una esfera tiene curvatura positiva, un plano tiene curvatura cero, y un espacio con forma de silla de montar tiene curvatura negativa.

Dimensión. El número de coordenadas requerido para especificar la posición de un punto en un espacio dado. Por ejemplo el plano tiene dimensión 2 y el espacio en el que vivimos (tal como lo modela la geometría de Euclides) tiene dimensión 3.

Disco (topológico). Una región en una superficie que puede ser deformada continuamente para dar un círculo más su interior.

Dodecaedro. Un sólido cuyas caras son 12 pentágonos regulares. (Véase Figura 38).

Dodecaedro rómbico. Un sólido cuya frontera se compone de 12 rombos idénticos - paralelogramos con todos los lados iguales. (Véase Figura 15).

Ecuación cuadrática. Cualquier ecuación $ax^2 + bx + c = 0$ donde x es una incógnita y a, b, c son constantes.

Ecuación cúbica. Cualquier ecuación $ax^3 + bx^2 + cx + d = 0$ donde x es una incógnita y a, b, c son constantes.

Ecuación diferencial. Una ecuación que relaciona una función con su ritmo de cambio.

Ecuación diofántica. Una ecuación para la que se exige que las soluciones sean números racionales.

Ecuación en derivadas parciales. Una ecuación diferencial que incluye los ritmos de cambio de una función con respecto a dos o más variables diferentes (a menudo espacio y tiempo).

Eje de rotación. Una recta fija alrededor de la cual rota un objeto.

Empaquetamiento. Una colección de formas dispuestas en el espacio de modo que no se solapan.

Empaquetamiento reticular. Una colección de círculos o esferas idénticos cuyos centros forman un retículo.

Entero. Cualquiera de los números..., -3, -3, -1, 0, 1, 2, 3...

Entero algebraico: Un número complejo que satisface una ecuación polinómica con coeficientes enteros y máximo coeficiente 1. Por ejemplo $i\sqrt{2}$, que satisface la ecuación $x^2 + 2 = 0$.

Entero/número ciclotómico. Una suma de potencias de una raíz compleja de la unidad con coeficientes enteros/racionales.

Esfera. El conjunto de todos los puntos del espacio a una distancia dada de un punto fijo, el centro. Es redonda, como una bola, pero el término «esfera» se refiere solo a los puntos de la superficie de la bola, no al interior.

3-esfera. Análogo tridimensional de una esfera: el conjunto de todos los puntos en un espacio tetradimensional a una distancia dada de un punto fijo, el centro.

Espacio topológico. Una forma que se considera que es «la misma» si se somete a cualquier transformación continua.

Estable. Un estado de un sistema dinámico al que el sistema vuelve si es sometido a una pequeña perturbación.

Exponente. En una potencia de una variable x, el exponente es la potencia en cuestión. Por ejemplo, en x^7 el exponente es 7.

Factorización. El proceso que escribe un número en términos de sus divisores primos. Por ejemplo, la factorización de 60 en primos es $2^2 \times 3 \times 5$.

Factorización en primos única. La propiedad de que cualquier número puede escribirse como un producto de números primos de una sola manera, exceptuando el cambio del orden en que se escriben los factores. Esta propiedad es válida para enteros, pero puede no serlo en sistemas algebraicos más generales.

Fase. Un número complejo sobre el círculo unidad, utilizado para multiplicar una función de onda cuántica.

Flujo de Ricci. Una ecuación que prescribe cómo cambia en el tiempo la curvatura de un espacio.

Frontera. El contorno de una región especificada.

Función. Una regla *f* que, cuando se aplica a un número *x*, produce otro número *f* (*x*). Por ejemplo, si *f* (*x*) = log *x* entonces *f* es la función logarítmica. La variable *x* puede ser real o compleja (en cuyo caso suele escribirse como *z*). Con más generalidad, *x* y *f* (*x*) pueden ser miembros de conjuntos específicos; en particular, el plano o el espacio.

Función de onda cuántica. Una función matemática que determina las propiedades de un sistema cuántico.

Función elíptica. Una función compleja que permanece invariable cuando dos números complejos independientes se suman a su variable. Es decir, *f* (*z*) = *f* (*z*+*u*) = *f* (*z*+*v*) donde *v* no es un múltiplo real de *u*. (Véase Figura 30).

Función zeta. Una función compleja introducida por Riemann que representa analíticamente los números primos. Está definida por la serie:

que converge cuando la parte real de *s* es mayor que 1. Esta definición puede extenderse a todo complejo *s*, excepto 1, mediante un proceso llamado prolongación analítica.

Género. El número de agujeros en una superficie.

Geometría no euclídea. Una alternativa a la geometría de Euclides en la que siguen siendo válidas todas las propiedades habituales de puntos y líneas, excepto la existencia de una única recta paralela a

una recta dada y que pasa por un punto dado. Hay dos tipos: elíptica e hiperbólica.

Geometría proyectiva. Un tipo de geometría en la que no existen rectas paralelas: dos rectas cualesquiera se juntan en un único punto. Se obtiene a partir de la geometría euclídea añadiendo una nueva «recta en el infinito».

Grado. La máxima potencia de la variable que aparece en un polinomio. Por ejemplo el grado de $6x^3 - 5x^2 + 4x - 7$ es 3.

Grupo. Una estructura algebraica abstracta que comprende un conjunto y una regla para combinar dos elementos cualesquiera del conjunto, sujeta a tres condiciones: la ley asociativa, la existencia de un elemento identidad y la existencia de inversos.

Grupo de cohomología. Una estructura algebraica abstracta asociada con un espacio topológico, análoga pero «dual» del grupo de homología.

Grupo fundamental. El grupo formado por clases de homotopía de lazos en un espacio topológico, bajo la operación «viajar a lo largo del primer lazo y luego a lo largo del segundo».

Grupo trivial. Un grupo que consiste solo en un único elemento, la identidad.

Homología (grupo). Un invariante topológico de un espacio, definido por lazos cerrados. Dos de tales lazos son homólogos si su diferencia es la frontera de un disco topológico.

Homotopía (grupo). Un invariante topológico de un espacio, definido por lazos cerrados. Dos de tales lazos son homotópicos si

cada uno de ellos puede ser deformado de forma continua hasta dar el otro.

Ideal (número). Un número que no está contenido en un sistema dado de números algebraicos, pero está relacionado con dicho sistema de una manera que recupera la factorización en primos única en casos en que la propiedad no se cumple. Reemplazado en el álgebra moderna por un ideal, que es un tipo especial de subconjunto del sistema concernido.

Ideal primo. Un análogo de un número primo para sistemas de números algebraicos.

Índice. El número de veces que una curva gira en sentido contrario a las agujas del reloj en torno a un punto escogido.

Inducción. Un método general para demostrar teoremas sobre números enteros. Si una propiedad es válida para 0, y su validez para cualquier número entero n implica su validez para $n + 1$, entonces la propiedad es válida para todos los números enteros.

Inestable. Un estado de un sistema dinámico al que el sistema no puede volver si es sometido a una pequeña perturbación.

Integral. Una operación del cálculo infinitesimal, que en efecto suma muchas pequeñas contribuciones. La integral de una función es el área contenida bajo su gráfica.

Integral logarítmica. La función $\mathrm{Li}(x) =$

Lazo. Una curva cerrada en un espacio topológico.

L-función de Dirichlet. Una generalización de la función zeta de Riemann.

Logaritmo. El logaritmo (natural) de *x*, escrito log *x*, es la potencia a la que hay que elevar e (= 2,71828...) para obtener *x*. Es decir, $e^{\log x} = x$.

Máximo. El valor más grande de algo.

Mínimo. El valor más pequeño de algo.

Modelo Estándar. Un modelo mecano cuántico que da cuenta de todas las partículas fundamentales conocidas.

Momento lineal. Masa multiplicada por velocidad.

Momento angular. Una medida de la cantidad de rotación que tiene un cuerpo.

NP-completo. Un problema específico de clase NP, con la propiedad de que si existe un algoritmo de clase P para resolverlo, entonces cualquier problema NP puede resolverse utilizando un algoritmo de clase P.

Número algebraico. Un número complejo que satisface una ecuación polinómica con coeficientes enteros, o equivalentemente coeficientes racionales. Por ejemplo $i\sqrt{2}/3$, que satisface la ecuación $x^2 + 2/9 = 0$, o equivalentemente $9x^2 + 2 = 0$.

Número complejo. Un número de la forma $a + b$ i donde i es $\sqrt{-1}$ y a, b son números reales.

Número compuesto. Un número entero que puede obtenerse multiplicando dos números enteros más pequeños.

Número congruente. Un número que puede ser la diferencia común de una secuencia de tres cuadrados de números racionales.

Número de Fermat. Un número de la forma $2^{2^k} + 1$ donde k es un número natural. Si este número es primo entonces se denomina *primo de Fermat*.

Número irracional. Un número real que no es racional; es decir, no es de la forma p/q donde p y q son enteros y $q \neq 0$. Ejemplos son $\sqrt{2}$ y π.

Número natural. Cualquiera de los números 0, 1, 2, 3...

Número primo. Un número entero mayor que 1 que no puede obtenerse multiplicando dos números enteros más pequeños. Los primeros números primos son 2, 3, 5, 7, 11, 13.

Número racional. Un número real de la forma p/q donde p y q son enteros y $q \neq 0$. Un ejemplo es 22/7.

Número real. Cualquier número que puede expresarse en forma decimal, que posiblemente continúa de forma indefinida; por ejemplo, π = 3,14159265358979323 85...

Número trascendente. Un número que no satisface ninguna ecuación algebraica con coeficientes racionales. Ejemplos son π y e.

Onda. Una perturbación que se propaga a través de un medio tal como un sólido, un líquido o un gas, sin producir un cambio permanente en el medio.

Operador. Un tipo especial de función A, que cuando se aplica a un vector v da otro vector Av. Debe satisfacer las condiciones de linealidad $A(v + w) = Av + Aw$ y $A(av) = aA(v)$ para cualquier constante a.

Optimización. Encontrar el máximo o el mínimo de una función.

Partícula. Una masa concentrada en un punto.

Pentágono. Un polígono de cinco lados.

Periódico. Algo que repite el mismo comportamiento indefinidamente.

Poliedro. Un sólido cuya frontera consiste en un número finito de polígonos.

Polígono. Una forma plana cuya frontera consiste en un número finito de líneas rectas.

Polígono regular. Un polígono cuyos lados tienen todos la misma longitud, y cuyos ángulos son todos iguales. (Véase Figura 4.)

Polinomio. Una expresión algebraica como $6x^3 - 5x^2 + 4x - 7$, en la que potencias de una variable x están multiplicadas por constantes y sumadas.

Polinomio irreducible. Un polinomio que no puede obtenerse multiplicando dos polinomios de grado menor.

Potencia. Un número multiplicado por sí mismo un número especificado de veces. Por ejemplo, la potencia cuarta de 3 es $3\times3\times3\times3 = 81$, simbolizada como 3^4.

Raíz de la unidad. Un número complejo ζ para el que alguna potencia ζ^k es 1. (Véase Figura 7 y nota 53).

Rango. El máximo número de soluciones racionales independientes de la ecuación que define una curva elíptica. «Independientes» significa que no pueden deducirse de otras soluciones utilizando una construcción geométrica estándar que combine dos soluciones cualesquiera para dar una tercera. (Véase Figura 25).

Razón. La razón de dos números a y b es a/b.

Red. Un conjunto de puntos (nodos) unidos por líneas (aristas).

Red cúbica de caras centradas. Un conjunto de puntos que se repiten en el espacio, obtenidos apilando cubos como un tablero de ajedrez tridimensional, y luego tomando las esquinas de los cubos y los centros de sus seis caras cuadradas. (Veánse Figuras 17, 19).

Red dual. Una red obtenida a partir de una red dada asociando un punto con cada región y uniendo puntos por aristas si las regiones correspondientes son adyacentes. (Véase Figura 10).

Relatividad general. Teoría de la gravitación de Einstein, en la que la fuerza de la gravedad se interpreta como la curvatura del espaciotiempo.

Retículo. En el plano: un conjunto de puntos que repite su forma a lo largo de dos direcciones independientes, como pautas de papel de pared (véase Figura 26). En el espacio: un conjunto de puntos que repite su forma a lo largo de tres direcciones independientes, como los átomos en un cristal.

Rotación. En el plano: una transformación en la que todos los puntos se mueven un mismo ángulo alrededor de un centro dado. En el espacio: una transformación en la que todos los puntos se mueven un mismo ángulo alrededor de una recta dada, el eje.

Secuencia. Una lista de números dispuestos en orden. Por ejemplo, la secuencia 1, 2, 4, 8, 16... de potencias de 2.

Secuencia aritmética. Una secuencia de números en la que cada número sucesivo es el anterior más una cantidad fija, la diferencia común. Por ejemplo, 2, 5, 8, 11, 14... con diferencia común 3. El término más antiguo es «progresión aritmética».

Serie. Una expresión en la que se suman muchas cantidades, a menudo infinitas.

Serie de potencias. Como un polinomio salvo que pueden ocurrir infinitas potencias de la variable; por ejemplo, $1 + 2\,x + 3\,x^2 + 4\,x^3 + \ldots$ En circunstancias apropiadas a esta suma infinita se le puede asignar un valor bien definido, y se dice que la serie converge.

Seno. Una función trigonométrica de un ángulo, definida por seno $A = b/c$ en la Figura 51.

Simetría. Una transformación de un objeto que deja invariable su forma global. Por ejemplo, rotar un cuadrado un ángulo recto.

Simetría gauge. Un grupo de simetrías locales de un sistema de ecuaciones: transformaciones de las variables que pueden variar de un punto a otro en el espacio, con la propiedad de que cualquier solución de las ecuaciones sigue siendo solución con tal de que en las ecuaciones se haga un cambio compensatorio con una interpretación física razonable.

Singularidad. Un punto en el que sucede algo desagradable, tal como que una función se haga infinita o deje de existir una solución de una ecuación.

Sistema dinámico. Cualquier sistema que cambia con el tiempo de acuerdo con reglas específicas. Por ejemplo, el movimiento de los planetas en el Sistema Solar.

Sólido regular. Un sólido cuya frontera está compuesta de polígonos regulares idénticos, dispuestos de la misma manera en cada vértice. Euclides demostró que existen exactamente cinco sólidos regulares.

Superficie. Una forma en el espacio obtenida uniendo regiones que son topológicamente equivalentes al interior de un círculo. Ejemplos son la esfera y el toro.

Tangente. Una función trigonométrica de un ángulo, definida por $\tan A = b/a$ en la Figura 51.

Teoría cuántica de campos. Una teoría mecanocuántica de una cantidad que llena el espacio y puede (y normalmente lo hace) tener valores diferentes en localizaciones diferentes.

Teoría gauge. Una teoría cuántica de campos con un grupo de simetrías gauge.

Tiempo de explosión. El tiempo más allá del cual deja de existir una solución de una ecuación diferencial.

Topología. El estudio de espacios topológicos.

Toro. Una superficie como la de un donut con un agujero. (Véase Figura 12).

Toro plano. Un toro obtenido identificando lados opuestos de un cuadrado, cuya geometría natural tiene curvatura cero. (Véase Figura 12).

Transformación. Otra palabra para «función», comúnmente utilizada cuando las variables implicadas son puntos en un espacio. Por ejemplo, «rotar alrededor del centro un ángulo recto» es una transformación de un cuadrado.

Transformación continua. Una transformación de un espacio con la propiedad de que puntos que están muy próximos no se separan mucho.

Traslación. Una transformación del espacio en la que todos los puntos se desplazan la misma distancia y en la misma dirección.

Triangulación. División de una superficie en una red de triángulos, o su análogo multidimensional.

Tripleta pitagórica. Tres números enteros, a, b, c tales que $a^2 + b^2 = c^2$. Por ejemplo, $a = 3$, $b = 4$, $c = 5$. Por el teorema de Pitágoras, números de este tipo forman los lados de un triángulo rectángulo.

Trisección. Dividir en tres partes iguales, especialmente en relación con ángulos.

Variable. Una cantidad que puede tomar cualquier valor en cierto rango.

Variedad. Una forma en el espacio definida por un sistema de ecuaciones polinómicas.

Variedad. Un análogo multidimensional de una superficie suave.

Variedad algebraica. Un espacio multidimensional definido por un conjunto de ecuaciones algebraicas.

Vector. En mecánica, una cantidad con magnitud y dirección. En álgebra y análisis, una generalización de esta idea.

Velocidad. El ritmo al que cambia la posición con respecto al tiempo. La velocidad tiene una magnitud y una dirección.

Vórtice. Fluido que fluye dando vueltas como un remolino. Puede tener cualquier tamaño, incluso muy pequeño.

Lecturas adicionales

Los libros marcados "*" son técnicos.

* Adams, Colin C., *The Know Book*, W. H. Freeman, 1994.

* Browder, Felix (ed.), « *Mathematical Developments Arising from Hilbert Problems* » (2 vols.), en *Proceedings of Symposia in Pure Mathematicas* 28, American Mathematical Society, 1976.

* Cao, Tian Yu, *Conceptual Developments of 20th Century Field Theories*, Cambridge University Press, 1997.

Cook, William J., *In Pursuit of the Travelling Salesman*, Princeton University Press, 2012.

Devlin, Keith, *The Millennium Problems*, Granta, 2004.

Diacu, Florin y Philip Holmes, *Celestial Encounters*, Princeton University Press, 1999.

Dudley, Underwood, *A Budget of Trisections*, Springer, 1987.

—, *Mathematical Cranks*, Mathematical Association of America, 1992.

Du Sautoy, Marcus, *The Music of the Primes*, Harper Perennial, 2004. [Hay traducción en castellano: *La música de los números primos*, Acantilado, Barcelona, 2013].

Gessen, Masha, *Perfect Rigour*, Houghton Mifflin, 2009.

* Goldman, Jay R., *The Queen of Mathematics*, A. K. Peters, 1998.

Hadamard, Jacques, *The Psychology of Invention in the Mathematical Field*, Dover, 1954.

* Hancock, Harris, *Lectures on the Theory of Elliptic Functions*, Dover, 1958.

Kaku, Michio, *Hyperspace*, Oxford University Press, 1994. [Hay traducción en castellano: *Hiperespacio*, Crítica, Barcelona, 2007].

Lagarias, Jeffrey C., *The Ultimate Challenge: The 3x + 1 Problem*, American Mathematical Society, 2011.

Livingston, Charles, « *Knot Theory* », Carus Mathematical Monographs, n.º 24, Mathematical Association of America, 1993.

Livio, Mario, *The Equation That Coludn't Be Solved*, Simon and Schuster, 2005. [Hay traducción en castellano: *La ecuación jamás resuelta*, Ariel, Barcelona, 2007].

McKean, Henry y Victor Moll, *Elliptic Curves*, Cambridge University Press, 1997.

O'Shea, Donald, *The Poincaré Conjecture*, Walker, 2007. [Hay traducción en castellano: *La conjetura de Poincaré*, Tusquets, Barcelona, 2008].

Randall, Lisa, *Warped Passages*, Allen Lane, 2005. [Hay traducción en castellano: *Universos ocultos*, Acantilado, Barcelona, 2013].

Ringel, Gerhard, *Map Color Theorem*, Springer, 1974.

Rogers, C. Ambrose, *Packing and Covering*, Cambridge Tracts in Mathematics and Mathematical Physics, n.º 54, Cambridge University Press, 1964.

Sabbagh, Karl, *Dr Riemann's Zeros*, Atlantic Books, 2002.

Sample, Ian, *Massive*, Basic Books, 2010.

Schoof, René, *Catalan's Conjecture*, Springer, 2008.

Singh, Simon, *Fermat's Last Theorem*, Fourth Estate, 1997. [Hay traducción en castellano: *El enigma de Fermat*, Editorial Planeta, Barcelona, 2003].

Stewart, Ian, *From Here to Infinity*, Oxford University Press, 1996. [Hay traducción en castellano: *De aquí al infinito*, Crítica, Barcelona, 2004].

—, *Why Beauty is Truth*, Basic Books, 2007. [Hay traducción en castellano: *Belleza y verdad*, Crítica, Barcelona, 2008].

—, *Seventeen Equations that Changed the World*, Profile, 2012. [Hay traducción en castellano: *17 ecuaciones que cambiaron el mundo*, Crítica, Barcelona, 2013].

Szpiro, George, *Kepler's Conjecture*, Wiley, 2003.

Tignol, Jean-Pierre, *Galois' Theory of Algebraic Equations*, Longman Scientific and Technical, 1980.

Watkins, Matthew, *The Mystery of the Prime Numbers*, Inamorata Press, 2010.

Wilson, Robin, *Four Colours Suffice*, Allen Lane, 2002.

Yandell, Benjamin, *The Honors Class*, A. K. Peters, 2002.

---

[i] Se trataba del cálculo de la estructura del grupo de Lie E_8». (*N. del t.*).

[ii] Este congreso tuvo lugar en Madrid. (*N. del t.*).

[iii] A veces se traduce «gauge» por calibre o aforo. Un término castellano similar es galga. No obstante, el término «gauge» ya está asentado en el lenguaje científico. (*N. del t.*).

[iv] En castellano no existe esta distinción. La misma palabra «variedad» traduce los términos ingleses «variety» y «manifold». (*N. del t.*).

[v] En el caso del alfabeto español, de 27 letras, el número de maneras es:

$$27! = 293.999.475.161.295.508.340.736.000.000$$