The Millennium Prize Problems

J. Carlson, A. Jaffe, and A. Wiles, Editors

Published by the **Clay Mathematics Institute**, Cambridge, Massachusetts,
jointly with the **American Mathematical Society**, Providence, Rhode Island

A list of picture credits is included at the end of the volume.

2000 *Mathematics Subject Classification.* Primary 00Bxx; Secondary 01Axx, 11G05, 14C30, 35Q30, 57R60, 03D15, 11M26, 81T13.

For additional information and updates on this book, visit
www.ams.org/bookpages/mprize-clay

Library of Congress Cataloging-in-Publication Data

The millennium prize problems.
Providence, RI : American Mathematical Society, 2006.
p. cm.
ISBN 0-8218-3679-X

20060444570

Published by the American Mathematical Society, Providence, RI,
for The Clay Mathematics Institute, Cambridge, MA.
Printed in the United States of America.

(∞) The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at http://www.ams.org/
Visit The Clay Mathematics Institute home page at http://www.claymath.org/

10 9 8 7 6 5 4 3 2 1 11 10 09 08 07 06

The Millennium Prize Problems

Contents

Introduction vii

Landon T. Clay xi

Statement of the Directors and the Scientific Advisory Board xv

A History of Prizes in Mathematics
Jeremy Gray 3

The Birch and Swinnerton-Dyer Conjecture
Andrew Wiles 31

The Hodge Conjecture
Pierre Deligne 45

Existence and Smoothness of the Navier–Stokes Equation
Charles L. Fefferman 57

The Poincaré Conjecture
John Milnor 71

The P versus NP Problem
Stephen Cook 87

The Riemann Hypothesis
Enrico Bombieri 107

Quantum Yang–Mills Theory
Arthur Jaffe and Edward Witten 129

Rules for the Millennium Prizes 153

Authors' Biographies 157

Picture Credits 161

Introduction

The Clay Mathematics Institute (CMI) grew out of the longstanding belief of its founder, Mr. Landon T. Clay, in the value of mathematical knowledge and its centrality to human progress, culture, and intellectual life. Discussions over some years with Professor Arthur Jaffe helped shape Mr. Clay's ideas of how the advancement of mathematics could best be supported. These discussions resulted in the incorporation of the Institute on September 25, 1998, under Professor Jaffe's leadership. The primary objectives and purposes of the Clay Mathematics Institute are "to increase and disseminate mathematical knowledge; to educate mathematicians and other scientists about new discoveries in the field of mathematics; to encourage gifted students to pursue mathematical careers; and to recognize extraordinary achievements and advances in mathematical research." CMI seeks to "further the beauty, power and universality of mathematical thinking."

Very early on, the Institute, led by its founding scientific board — Alain Connes, Arthur Jaffe, Edward Witten, and Andrew Wiles — decided to establish a small set of prize problems. The aim was not to define new challenges, as Hilbert had done a century earlier when he announced his list of twenty-three problems at the International Congress of Mathematicians in Paris in the summer of 1900. Rather, it was to record some of the most difficult issues with which mathematicians were struggling at the turn of the second millennium; to recognize achievement in mathematics of historical dimension; to elevate in the consciousness of the general public the fact that, in mathematics, the frontier is still open and abounds in important unsolved problems; and to emphasize the importance of working toward solutions of the deepest, most difficult problems.

After consulting with leading members of the mathematical community, a final list of seven problems was agreed upon: the Birch and Swinnerton-Dyer Conjecture, the Hodge Conjecture, the Existence and Uniqueness Problem for the Navier–Stokes Equations, the Poincaré Conjecture, the **P** versus **NP** problem, the Riemann Hypothesis, and the Mass Gap problem for Quantum Yang–Mills Theory. A set of rules was established, and a prize fund of US$7 million was set up, this sum to be allocated in equal parts to the seven problems. No time limit exists for their solution.

The prize was announced at a meeting on May 24, 2000, at the Collège de France. On page xv we reproduce the original statement of the Directors and the Scientific Advisory Board. John Tate and Michael Atiyah each spoke about the Millennium Prize Problems: Tate on the Riemann Hypothesis, the Birch and Swinnerton-Dyer Problem, and the **P** vs **NP** problem; Atiyah on the Existence and Uniqueness Problem for the Navier–Stokes Equations, the Poincaré Conjecture, and the Mass Gap problem for Quantum Yang–Mills Theory. In addition, Timothy Gowers gave a public lecture, "On the Importance of Mathematics". The lectures — audio, video, and slides — can be found on the CMI website: `www.claymath.org/millennium`.

The present volume sets forth the official description of each of the seven problems and the rules governing the prizes. It also contains an essay by Jeremy Gray on the history of prize problems in mathematics.

The editors gratefully acknowledge the work of Candace Bott (editorial and project management), Sharon Donahue (photo and photo credit research), and Alexander Retakh (TEX, technical, and photo editor) for their care and expert craftsmanship in the preparation of this manuscript.

James Carlson, Arthur Jaffe, and Andrew Wiles

Landon T. Clay

Founder

Clay Mathematics Institute

Landon T. Clay

Landon T. Clay

Landon T. Clay has played a leadership role in a variety of business, science, cultural, and philanthropic activities. With his wife, Lavinia D. Clay, he founded the Clay Mathematics Institute and has served as its only Chairman. His past charitable activities include acting as Overseer of Harvard College, as a member of the National Board of the Smithsonian Institute, and as Trustee of the Middlesex School. He is currently a Great Benefactor and Trustee Emeritus of the Museum of Fine Arts in Boston and for 30 years has been Chairman of the Caribbean Conservation Corporation, which operates a turtle nesting station in Costa Rica. He donated the Clay Telescope to the Magellan program of Harvard College in Chile. The Clay family built the Clay Science Centers at Dexter School and Middlesex School. He received an A.B. in English, cum laude, from Harvard College.

Board of Directors
and
Scientific Advisory Board

Board of Directors and Scientific Advisory Board
Landon T. Clay, Lavinia D. Clay, Finn M.W. Caspersen,
Alain Connes, Edward Witten, Andrew Wiles, Arthur Jaffe
(not present: Randolph R. Hearst III and David R. Stone)

Statement of the Directors and the Scientific Advisory Board

In order to celebrate mathematics in the new millennium, the Clay Mathematics Institute of Cambridge, Massachusetts, has named seven "Millennium Prize Problems". The Scientific Advisory Board of CMI selected these problems, focusing on important classic questions that have resisted solution over the years. The Board of Directors of CMI have designated a US$7 million prize fund for the solution to these problems, with US$1 million allocated to each. During the Millennium meeting held on May 24, 2000, at the Collège de France, Timothy Gowers presented a lecture entitled "The Importance of Mathematics", aimed for the general public, while John Tate and Michael Atiyah spoke on the problems. CMI invited specialists to formulate each problem.

One hundred years earlier, on August 8, 1900, David Hilbert delivered his famous lecture about open mathematical problems at the second International Congress of Mathematicians in Paris. This influenced our decision to announce the millennium problems as the central theme of a Paris meeting.

The rules that follow for the award of the prize have the endorsement of the CMI Scientific Advisory Board and the approval of the Directors. The members of these boards have the responsibility to preserve the nature, the integrity, and the spirit of this prize.

Directors: Finn M.W. Caspersen, Landon T. Clay, Lavinia D. Clay, Randolph R. Hearst III, Arthur Jaffe, and David R. Stone

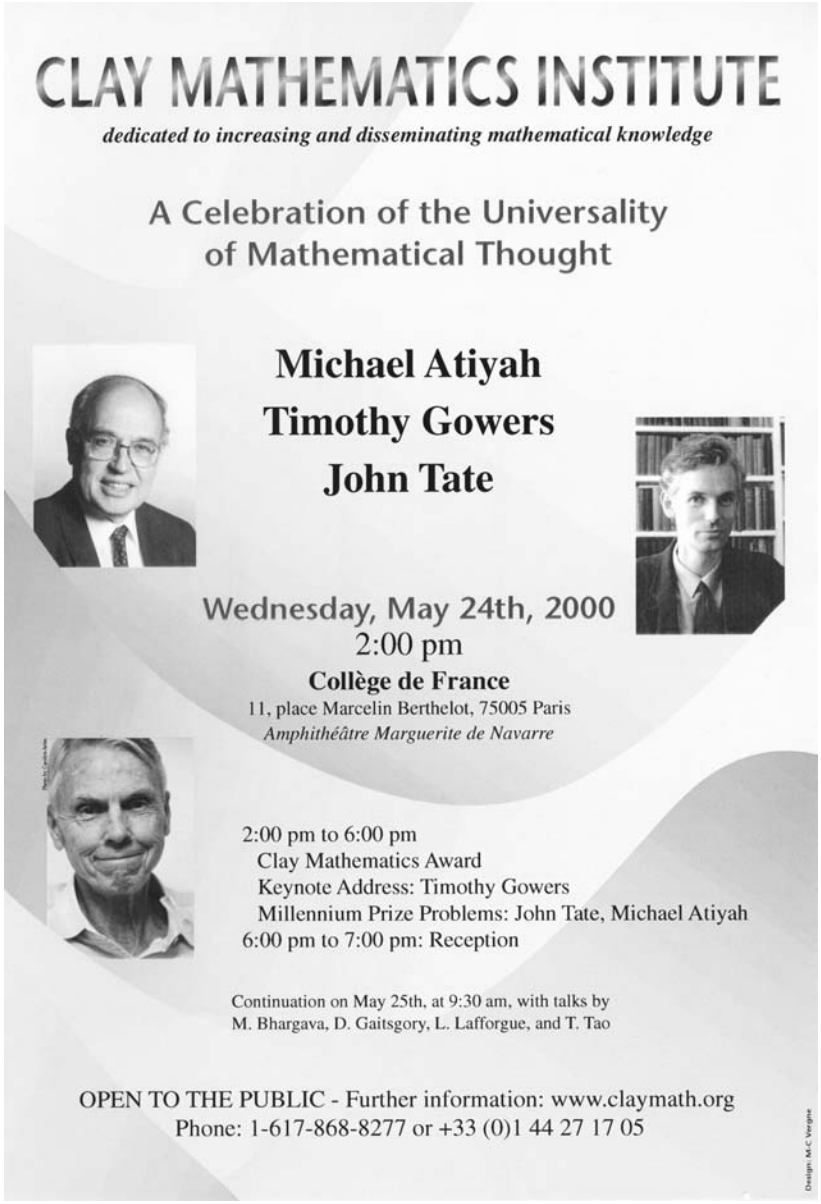
Scientific Advisory Board: Alain Connes, Arthur Jaffe, Andrew Wiles, and Edward Witten

Paris, May 24, 2000

Collège de France

Paris meeting

Alain Connes, Collège de France, and David Ellwood, Clay Mathematics Institute, undertook the planning and organization of the Paris meeting, assisted by the generous help of the Collège de France and the CMI staff. The videos of the meeting, available at `www.claymath.org/millennium`, were shot and edited by François Tisseyre.



Poster of the Paris meeting

A History of Prizes in Mathematics

Jeremy Gray

A History of Prizes in Mathematics

Jeremy Gray

1. Introduction

Problems have long been regarded as the life of mathematics. A good problem focuses attention on something mathematicians would like to know but presently do not. This missing knowledge might be eminently practical, it might be wanted entirely for its own sake, its absence might signal a weakness of existing theory — there are many reasons for posing problems. A good problem is one that defies existing methods, for reasons that may or may not be clear, but whose solution promises a real advance in our knowledge.

In this respect the famous three classical problems of Greek mathematics are exemplary. The first of these asks for the construction of a cube twice the volume of a given cube. The second asks for a method of trisecting any given angle, and the third for the construction of a square equal in area to a given circle.[1] Because Euclid, in his *Elements*, used only straight edge and circle (ruler and compass) to construct figures, a modern interpretation of the problems has restricted the allowed solution methods to ruler and compass constructions, but none of the Greek attempts that have survived on any of these problems obey such a restriction, and, indeed, none of the problems can be solved by ruler and compass alone. Instead, solutions of various kinds were proposed, involving ingenious curves and novel construction methods, and there was considerable discussion about the validity of the methods that were used. A number of distinguished mathematicians joined in, Archimedes among them, and it seems that the problems focused attention markedly on significant challenges in mathematics.

In addition to the contributions to mathematics that the problems elicited, there is every sign that they caught the public's attention and were regarded as important. Socrates, in Plato's dialogue *Meno*, had drawn out

[1] To speak of just classical problems is something of a misnomer. There were other equally important problems in classical times, such as the construction of a regular seven-sided polygon.

of a slave boy the knowledge of how to construct a square twice the size of a given square, thus demonstrating his theory of knowledge. Plato claimed that the analogous problem of duplicating the cube was ordained by the Gods, who required the altar at Delos to be doubled exactly. Less exaltedly, the problem of squaring the circle rapidly became a by-word for impossibility, and Aristophanes, a contemporary of Plato's, could get a laugh from an Athenian audience by introducing a character who claimed to have done it. Since all these problems possess simple, approximate, 'engineering' solutions, the Greek insistence on exact, mathematically correct, solutions is most striking.

To solve an outstanding problem is to win lasting recognition, as with the celebrated solution of the cubic equation by numerous Italian mathematicians at the start of the 16th century. In 1535, Tartaglia was challenged by one Antonio Fior to solve 30 problems involving a certain type of cubic equation. Fior had been taught the solution to the cubic by Scipione del Ferro of Bologna, who seems to have discovered it. As was the custom of the day, Tartaglia replied with 30 problems of his own on other topics, two months in advance of the contest date. With one day to go, Tartaglia discovered the solution method for Fior's cubics and won the contest and the prize, which was thirty dinners to be enjoyed by him and his friends. Such contests naturally promoted secrecy rather than open publication, because only the solutions but not the methods had to be revealed. Tartaglia later divulged the method in secret to Cardano, who some years later published it in his *Ars Magna* in 1545. Cardano argued that since the original discovery was not Tartaglia's, he had had no right ask that it be kept secret. Moreover, by then Cardano had extended the solution to all types of cubic equations, and his student, Ferrari, had gone on to solve the quartic equation as well.[2]

The tradition of setting challenging problems for one's fellow (or, perhaps, rival) mathematicians persisted. In 1697 the forceful Johann Bernoulli posed the brachistochrone problem, which asks for the curve joining two points along which a body will most quickly descend. He received three answers. Newton's he recognised at once: "I know the lion by his claw," he said. In fact, goaded by the way Bernoulli had wrapped the mathematical challenge up in the rapidly souring dispute over the discovery of the calculus, Newton had solved the problem overnight [**40**, p. 583].

Problems could be set to baffle rivals, but ultimately more credit resides with those who posed questions out of ignorance, guided by a shrewd sense of their importance. It is the lasting quality of the solution, a depth that brings out what was latent in the question, that is then recognised when the solver

[2]For some of the documents involved in this story, see [**18**, pp. 253–265]. Cubics were taken to be of different types because they were always taken with positive coefficients, so $x^3 + x = 6$ and $x^3 + 6 = x$ are of different types.

is remembered. Problems that point the way to significant achievements were systematically generated in the 18th century. This tradition was less successful in the 19th century, but was famously revived in a modified form by Hilbert in 1900. His choices of problems were often so inspired that those who solved one were said, by Hermann Weyl, to have entered the Honours Class of mathematicians [**41**]. It is this tradition of stimulating problems that the Clay Mathematics Institute has also sought to promote.

2. The Academic Prize Tradition in the 18th Century

The 18th century was the century of the learned academy, most notably those in Berlin, Paris, and St. Petersburg. To be called to one of these academies was the closest thing to a full-time research position available at the time, a chance to associate with other eminent and expert scholars, and the opportunity to pursue one's own interests. It was also a chance to influence the direction of research in a new and public way, by drawing attention to key problems and offering substantial rewards for solving them.

The academies ran their prize competitions along these lines. Problems would be set on specific topics. A fixed period of time, usually 18 months to two years, was allowed for their solution, a prize of either a medal or money was offered for their solution, and the solutions would usually be published in the academy's own journal. There was often a system of envelopes and mottos to assist anonymity, and success was liable to make one famous within the small world of the savants of the day. This was a group of some modest size, however, and was by no means confined to the very small group of mathematicians of the time. The historian Adolf Harnack (twin brother of the mathematician Axel) described the situation vividly in his history of the Berlin Academy of Sciences[3]:

> In a time when the energies and the organization for large scientific undertakings — with the exception of those in astronomy — were still lacking, the prize competitions announced annually by the academies in Europe became objects for scientific rivalries and the criterion for the standing and acumen of scientific societies.... This was so because specialities were most often disregarded and the themes chosen for competitions were either those that required perfect insight into the state of an entire discipline and its furtherance with respect to critical points, or those that posed a fundamental problem. The prize competitions constituted the lever by which the different sciences were raised one step higher

[3]This translation from [**13**, p. 12], original in [**24**, vol.1, pp. 396–397]. Reprinted with the permission of Cambridge University Press.

> from one year to another; in addition, they were important for universalizing and unifying science. The questions were addressed to learned men all over Europe and were communicated throughout the scientific world. The suspense surrounding the announcement of the question was, in fact, larger than that of the answer, for it was in the formulation of the question that mastery was revealed. The invitation was not addressed to young recruits of science but to the leaders who eagerly answered the call to contest. The foremost thinkers and learned men — Euler, Lagrange, d'Alembert, Condorcet, Kant, Rousseau and Herder — all entered the arena. This circumstance which may seem quite strange today requires special explanation. This latter ... resides in the fact that the learned man of the 18th century was still a *Universalphilosoph*. His mind could discern an abundance of problems in different scientific areas which all seemed equally attractive and enticing. Which one should he attack? At that moment, the Academy came to the rescue with its prize competitions. It presented him with a given theme and assured him a universally interested audience.

The first prize fund to be established was endowed by Count Jean Rouillé de Meslay, a wealthy lawyer, who left the Académie des Sciences in Paris 125,000 livres in his will in 1714 [**13**, p. 11].[4] The Académie took this up, and from 1719 on, prizes were to be awarded every two years. The first two topics concerned the movement of planets and celestial bodies and, a related issue at the time, the determination of longitude. These were substantial issues. Newton's novel theory of gravity, proclaimed in his *Principia*, was not widely accepted in Continental Europe. It sought to replace a clear physical process, vortices, with the much more problematic notion of action at a considerable distance, and it had a conspicuous flaw amid many striking successes: the motion of the moon. This particular failing was most unfortunate because the motion of the moon, if properly understood, could be a key to the longitude problem.

Daniel Bernoulli

[4]The standard source of information is [**32**]. It should be pointed out that 125,000 livres was a very large sum of money; a skilled artisan of the period might hope to earn 300 livres a year.

Among the more famous winners of the Paris academy prizes was Daniel Bernoulli, who won no less than ten prizes, and most of his contributions show how important the topic of navigation was. His first success came in 1725, for an essay on the best shape of hour-glasses filled with sand or water, such as might serve as nautical clocks. In 1734 he shared the prize with his father Johann, who begrudged him his success, for an essay exploring the effect of a solar atmosphere on planetary orbits. Later successes included a paper on the theory of magnetism (joint with his brother Johann II) and on the determination of position at sea when the horizon is not visible. He also wrote on such matters as how to improve pendulum clocks.

The Academy of Sciences in St. Petersburg was established on the orders of the Emperor Peter the Great on January 28 (February 8), 1724, and was officially opened in December 1725, shortly after his death. To ensure that it worked to the highest standards of the time, Peter hired several leading mathematicians and scientists, Euler, Nicholas and Daniel Bernoulli, and Christian Goldbach among them. Euler was only 20 when he arrived, and he remained associated with the Academy for most of his life, publishing in its journal prolifically even when he was not an Academician.

In Berlin, the rival Academy of Sciences, the Académie Royale des Sciences et de Belles Lettres de Berlin, was founded in 1700, but it did not become influential until it was reorganised along Parisian lines in 1743 by Frederick the Great, who had come to power in 1740 and reigned until his death in 1786. He wished the academy to be useful to the state, and he paid the new staff he brought in high salaries, more than they would get in Paris but less than St. Petersburg. He installed Maupertuis as director of the academy, and Euler as director of the mathematical class. Maupertuis supported Voltaire's turn toward the English: Newtonian mechanics and Lockean metaphysics as opposed to Cartesianism. The first prize topic, for 1745, was 'On electricity' and was won by Waitz, the Finance Minister in Kassel. The prize amounted to some 50 ducats, and from 1747 took the form of a gold medallion. In 1746 d'Alembert won the prize for his essay 'Réflexions sur la cause générale des vents', which was his response to the challenge: 'Determine the order and the law which the wind must follow if the Earth was entirely surrounded on all sides by ocean, in such a way that the direction and speed of the wind is determined at all times and for all places.'

Frederick the Great

Eleven entries had been submitted; d'Alembert's is the first in which partial differential equations were put to general use in physics [**39**, p. 96]. The famous wave equation appeared in a paper d'Alembert published in the Memoirs of the Berlin Academy the next year, 1747.

As further evidence of the interest generated by the Berlin prize competition, Harnack noted that there were often a dozen entries for a given problem, although it was generally impossible to know who entered because only the names of the winner (and sometimes a runner-up) were ever announced. Young and old could enter, and could enter successive competitions; there was an explicit rule that in the event of a tie the foreign competitor was to be preferred. In the course of the 18th century, twenty-six different winners were German, ten French, two Swiss, and one each came from Italy and Transylvania.

There was naturally some overlap between the academies [**24**, p. 398]. Some of the same names occur in the lists of the other academies, and some more than once, the most notable case being that of Euler, who won no less than twelve prizes from various academies.

Leonhard Euler

All of this work entailed continual involvement behind the scenes judging the essays. Decisions were final, but were not always accepted gracefully: d'Alembert in the early 1750s complained that he was the victim of a cabal in Berlin that had denied him a prize for an essay on fluid mechanics (in fact, no one won the prize that year). He thereupon published his own essay, in 1752, in which he raised the paradox that the flow round an elliptical object should be the same fore and aft, which implied that there would be no resistance to the flow. It was left to others to find the flaw in d'Alembert's argument, and meanwhile his relations with Euler worsened. The basic problem may have been one of temperament. D'Alembert, although a charming conversationalist, was a slow writer who did not express his ideas with clarity. Euler was unfailingly lucid and wrote with ease. D'Alembert may have come to resent the way in which his ideas, once published, were so readily taken up and well developed by the other man. It was only in 1764, when d'Alembert tried actively to intervene with Frederick the Great on Euler's behalf, that relations between Euler and d'Alembert were put on a more amicable footing. D'Alembert's interventions were unsuccessful, however, and Euler left Berlin permanently for St. Petersburg in 1766.

Over the years a few problems recurred, mostly to do with astronomy and navigation. Euler won the Paris Academy prize of 1748 for an investigation of the three-body problem (in this case Jupiter, Saturn, and the sun). Then, knowing that Clairaut was wrestling with the inverse square law and was prepared to modify it, Euler proposed the motion of the moon as a prize topic for the St. Petersburg Academy in 1751.

Clairaut rose to the challenge, and suddenly found that he need not abandon Newton's law, as he had at first thought, but that a different analysis of the problem showed that the law could indeed give the right results. His successful solution to this problem was one of the reasons that the inverse square law of gravity became established and other theories died out. Other reasons included Clairaut's successful prediction of the return of Halley's comet in 1759. Comets are, of course, particularly sensitive to the perturbative effect of the larger planets, so the challenge of determining their orbits highlighted the importance of the many-body problem in celestial mechanics, which the Berlin Academy returned to again, for example in 1774.

Alexis Claude Clairaut

The Paris Academy in 1764 asked for essays on the libration of the moon: Why does it always present more or less the same face to us, and what is the nature of its small oscillations? In 1765 they asked about the motion of the satellites of Jupiter, and the competition was won by Lagrange (who was then 29).

Both these topics reflect the hope that celestial motions could somehow be interpreted as clocks and so solve the longitude problem. In 1770 the prize went jointly to Euler and his son Albrecht for an essay on the three-body problem, and in 1772 the same topic again led to the prize being shared, this time between Euler and Lagrange. In 1774, Lagrange won again, for an essay on the secular motion of the moon, but he had begun to tire of the subject and needed an extension to the closing date, which d'Alembert requested Condorcet to offer as an inducement to continue. Lagrange refused to enter the next competition, on the motion of comets — the prize went to Nicholas Fuss — but he entered the competition on the same topic in 1780 and won the double prize of 4,000 livres. Thereafter he never entered a competition of the Paris Academy [**26**].

Prizes could be set to address embarrassing deficiencies in the state of the art. Lagrange, a member of the Berlin Academy since 1766, persuaded

it to ask for a rigorous foundation of the calculus in 1784. The prize was to be awarded for 'a clear and precise theory of what is called Infinity in mathematics'. The continual use of infinitely large and infinitely small quantities in higher mathematics, noted the preamble, was successful despite seeming to involve contradictions. What was needed was a new principle that would not be too difficult or tedious and should be presented 'with all possible rigour, clarity, and simplicity' [**20**, pp. 41–42]. The tedious approach the academy wished to head off was the defence of the Newtonian calculus that MacLaurin had mounted, which replaced Newton's intuitive limiting arguments with the cumbersome apparatus of double reduction ad absurdum.

Joseph Louis Lagrange

The competition was won by Simon L'Huillier, and two essays written for it made their way into books (L'Huillier's [**27**] and Lazare Carnot's [**8**]). The judges were satisfied with neither, however, and, when the newly founded École Polytechnique required Lagrange to publish his lectures he produced, his own account, the *Fonctions analytiques* of 1797. This entirely algebraic account lasted until Cauchy's analysis began to sweep it away in the 1820s.[5]

3. The Academic Prize Tradition in the 19th Century

After the French Revolution, the revised Académie in Paris had two new prizes, starting in 1803, of 3,000 FFr: the grand prix des sciences mathématiques or the Grand Prix des sciences physiques. The two were intended to alternate, and a professor's salary at the time was some 4,000 FFr, so the prize was indeed generous. There were also some irregular prizes, such as the competition proposed by Napoleon in 1809 on the vibrational modes of elastic plates. This was in response to Chladni, who had come to Paris and demonstrated many new experiments on this unstudied phenomenon. Laplace was in charge of the commission that was to judge the prize, and he hoped that it would provide an occasion to advance his protégé, Poisson. The prize competition was officially announced for 1811 and drew only one entry, not, however, from Poisson, but one written by the unknown Sophie Germain.[6] The judges found it inadequate, and the competition was extended to October 1813. Germain worked to deepen her analysis, and hers was again the only entry. She was by now in correspondence with Legendre,

[5]See [**9**]. The facsimile re-edition edited by U. Bottazzini has a very useful introduction.

[6]This account follows that in [**6**].

who was one of the judges, and he seems to have been disappointed with her work, although it now obtained an honourable mention. The competition was extended again, to October 1815. Only now did Poisson submit a memoir, but since he had been a judge of this very competition in 1813 his actions were irregular to say the least, and Legendre protested. The memoir was nonetheless read to the Institut de France and a note about it inserted in the *Correspondance de l'École Polytechnique*, where it was said that it might prove helpful to potential competitors. Poisson seems to have hoped that by acting in this way the question would be permanently withdrawn while he nonetheless earned the approval of Laplace. However, his actions were so scandalous that a deal seems to have been struck to keep the question open, and possibly even to give a prize to Sophie Germain if she could improve her work sufficiently. This in the end, she did, not mathematically, but experimentally, and she was awarded the prize in 1815.

Sophie Germain

As Germain's story shows, the administration of prizes in the small hot-house environment of Paris was not without problems. It may have discouraged Germain from entering the competition on Fermat's Last Theorem, which was the topic set in 1815 for 1817. In fact, no one entered, and after four years the question was withdrawn, but in that time Germain had made one of the few notable inroads on the topic in the century between Euler and Kummer [**17**, p. 62]. These results made her famous when Legendre published them as hers in his *Théorie des Nombres* in 1830.

Further evidence of the way the prize competition worked at the start of the 19th century is provided by the mathematics prize for 1815. This was won by Cauchy who, in answer to a question about the propagation of waves, wrote a memoir chiefly remarkable for his discovery of the way a

function and its Fourier transform are inter-related, made in ignorance of Fourier's own work. His memoir was not published, however, until 1827, by which time it had long been eclipsed by Poisson's independent discovery and prompt publication of many of the same results [**4**, p. 90].

A notable success of these prizes came in 1819, when the mathematics prize was won by Fresnel in a decisive moment for the wave theory of light. But hints of what was to come are visible in another celebrated award of the prize, which went jointly to Abel (posthumously) and Jacobi in 1830 for the independent discovery of elliptic functions, not because the topic had been identified in advance but because their work was rapidly recognised after the event to be a momentous discovery (and because Legendre was in a position to see that the Académie could offer such a prize). During the 19th century, the original tradition of what might be called prospective prizes (titles announced in advance and a specific deadline to be met) came under pressure, and the alternative of retrospective prizes or general subject area prizes for work in some area of mathematics or science were proposed. This was more and more the case as new prizes were established, but even when the title was precise the judges began to allow previous work to be submitted, rather than rolling the topic over for two more years (which was also done).

The difficulties that arose when a question attracted no good entries were considerable. It was embarrassing, and there were financial implications. The first hint of an alternative solution that preserved outward appearances had come in 1810, when Lagrange and Laplace jointly proposed the double refraction of light as the Grand Prix topic for 1810, knowing very well that the 35-year-old Etienne Malus was at that moment doing brilliant work in optics. He did indeed win, and happily the challenge inspired him to extend his earlier work considerably; during the process he discovered the polarisation of light [**14**, pp. 271–272]. This same thing happened in 1812, when heat diffusion was the topic, upon which Fourier was known to be at work, and this time the result was that great rarity, a work as important in the history of physics as it is in the history of mathematics.

As the 19th century wore on, the Grand Prix in mathematics had mixed results. A question on the perturbations of elliptic orbits was first set in 1840 but only answered successfully (by Hansen) in 1846, but other questions, on the maxima and minima of multiple integrals and on Abelian functions, were answered successfully within the initial two-year period, by Sarrus and Rosenhain respectively. Then the commissions ran out of luck for a while. Fermat's Last Theorem was proposed in 1850, with Cauchy as the chairman of the judges, but no satisfactory answers were received, and the problem was rolled over to 1853 before being abandoned. The spur for this was Lamé's argument using cyclotomic numbers, in 1847, in which he mistakenly

supposed that such integers have a unique factorisation law. His error was pointed out by Liouville, but this only inspired Cauchy to claim that he could solve the problem, and order was not restored until Liouville brought Kummer's much more profound ideas to France [**17, 30**]. The distribution of heat in an infinite body was the topic proposed in 1858 and finally withdrawn in 1868. This competition is remembered only because Riemann's entry was passed over — the jury found that the way in which the results had been discovered was insufficiently clear.

In 1865 Bertrand was in charge of a question asking for an improvement to the theory of second-order partial differential equations, but there were no answers, and the question was repeated; it was answered to the satisfaction of the panel by Bour in 1867. There were no answers to the question Bonnet set in 1867 on algebraic surfaces, and none to Puiseux's question (the three-body problem) in 1872. Singular moduli and complex multiplication in the theory of elliptic functions drew no response in 1874, nor did the suggestion that elliptic and Abelian functions might be profitably applied to the theory of algebraic curves in 1878. However, in between those years Darboux did win the prize for an essay on singular solutions of first-order partial differential equations.

Bernhard Riemann

In 1880 the Grand Prix was again awarded, for an essay 'significantly improving the theory of linear ordinary differential equations'. The prize went to G.H. Halphen, for an essay on the invariants associated to a differential equation, but the competition is best remembered for the second-place entry from Poincaré on the theory of automorphic functions and the relationship between non-Euclidean geometry and the nascent theory of Riemann surfaces (see [**35**] or [**22**] for fuller accounts).

In 1882 embarrassment came to the Paris academy. With Camille Jordan in charge, they proposed an investigation of the number of ways a number can be written as the sum of five squares. The young German mathematician Hermann Minkowski, then only 18, and the English mathematician H.J.S. Smith submitted entries that shared the prize. Unfortunately, Smith's contribution was confined to showing that he had already solved the problem some years before. To make matters worse, by the time the result was announced, Smith had died. Hostile critics pounced on this to suggest that Minkowski must have known of Smith's work because he was surely too young to have done the work on his own, and the ensuing row carried ugly

hints of anti-semitism before the academy rightly pronounced itself satisfied that Minkowski had been entirely independent of Smith [**15**].

They had better luck in 1886, when Halphen proposed a question generalising the regular solids that Goursat answered; in 1888 when Poincaré asked about algebraic equations in two independent variables, and Picard was awarded the prize; and in 1894 when Darboux asked for an improvement in the theory of the deformation of surfaces, and the commission was able to award the prize to Weingarten. A more famous award came in 1892 when Hermite persuaded the academy to ask for the determination of the number of prime numbers less than a given number (the prime number theorem), the aim being to draw mathematicians to fill some of the gaps in Riemann's famous paper of 1857. He hoped in this way to get his friend Stieltjes to write up the details in support of his 1885 claim to have solved the Riemann hypothesis. As the closing date drew near, and even though Hermite wrote to Stieltjes to encourage him, no essay was forthcoming. Instead, the young Hadamard presented his doctoral thesis on entire functions in 1890, and Hermite, who was one of the examiners, suggested that Hadamard find applications for his ideas. Hadamard confessed that he had none, but he soon realised that his new theory was just what was needed to resolve the prime number theorem, and he submitted a long essay to that effect, which was awarded the prize on 19 December 1892 [**33**, pp. 55–57]. Stieltjes never found the proof he had incautiously claimed.

These competitions continued after World War I, when Julia, Lattès, and Pincherle wrote essays on iteration theory [**1**, pp. 108–116]. The prize went to Julia, with an honourable mention and 2,000 FFr to Lattès. Fatou, who had decided not to enter the competition, was also awarded 2,000 FFr. His work and Julia's were strikingly similar, and at Julia's request the question of independence and priority was addressed directly. It was found that Julia's results, presented in a sealed letter to the academy as the competition rules required, did indeed predate Fatou's publications, but that the men had worked independently.

Regardless of the problems administering the prizes, donors found them attractive, until by 1850 there were thirteen different French prizes across the sciences, all controlled from Paris. The number rose again after the defeat of France in the Franco-Prussian War (1870–71), when there was a widespread feeling that science had been allowed to decline too far and thus contributed to the national defeat. The new prizes, like the more established ones, were overwhelmingly retrospective, but to make them more attractive it was argued that the reward should be financial rather than in the form of a medal. The impact of these prizes was considerable, amounting to one third to half of the winner's annual salary, depending on whether or not he or she

lived in Paris, and was often a huge boost to scientists when equipment was needed.

The strictly mathematical prizes participated in this general shift. The prix Poncelet was endowed by the wife of General Jean-Victor Poncelet after his death in 1868 in order to carry out his dying wish that the sciences be advanced. Poncelet himself had been one of the chief creators of projective geometry in the 1820s, before turning to the theory of machines. His widow's generosity was augmented by a further sum of money, and the prize of 2,000 francs was inaugurated in 1876. It operated invariably as a retrospective prize.

The prix Bordin was created by the will of Charles-Lauren Bordin, who died in 1835 leaving the Institut de France 12,000 francs. The institute eventually created an annual prize of 3,000 francs after the Company of Notaries had declined a similar bequest, and the first prize was awarded in 1854. Topics moved from theoretical physics to pure mathematics. In 1888, for example, the prize was offered for an essay improving in some important way the theory of motion of a solid body. There is good evidence [**11**] that the topic was set with Sonya Kovalevskaya's work in mind.

Sofia Kovalevskaya

Kovalevskaya wrote to Mittag-Leffler over the summer of 1888 to say that "Bertrand gave a large dinner in my honor, attended by Hermite, Picard, Halphen, and Darboux. Three toasts were proposed in my honor, and Hermite and Darboux said openly that they have no doubt that I shall have the prize" [**11**, p. 114]. Since Bertrand was the perpetual secretary of the academy, Hermite was the most influential French mathematician behind the scenes, and Darboux was on the panel of judges, they presumably knew whereof they spoke. They had already extended the closing date so that her essay could be received — it arrived late in the summer — and the prize was awarded to her in December. That said, she won the essay for a fine piece of analysis applying the theory of Abelian functions to rigid body motion, thus showing how the new functions had their uses in physical problems.

In 1892 the advertised topic was in differential geometry, and Gabriel Koenigs won with an essay on geodesics. In 1892 the topic was the application of the theory of Abelian functions to geometry, then the domain of a rising star, George Humbert, who duly won. In 1894 Paul Painlevé and

Roger Liouville shared the prize for their essays on the use of algebraic integrals in problems in mechanics. In 1896 Hadamard won with an essay on geodesics on surfaces, one of the few to respond to the work of Poincaré on the same subject. The feeling that all these topics were set with a shrewd eye to who was working actively on what subject deepens with what became one of the more awkward incidents in the history of the prize [**10**, pp. 58–59]. The first draft of the paper Enriques and Severi submitted to the committee of the Bordin prize was flawed. They became aware of these mistakes after a conversation with de Franchis, and withdrew their paper only to make some corrections and to re-submit it, even though they knew that Bagnera and de Franchis were also candidates for the same prize, and indeed had better results. The prize went to Enriques and Severi, and de Franchis complained through the intermediary of Guccia, the well-respected editor of the Palermo *Rendiconti* (Bagnera and de Franchis were also Sicilian). As a result, the same topic was advertised again as the prize for 1909, and this time Bagnera and de Franchis won the prize.

The early years of the Steiner prize from the University of Berlin illustrates the problems of prize competitions only too well. It was endowed in the will of the distinguished exponent of synthetic geometry, the Swiss mathematician Jacob Steiner, who had taught most of his life at Berlin University and died in 1863. Steiner stipulated that the prize, of 8,000 Thaler, be awarded once every two years for a geometric topic treated synthetically.[7] The first time the prize was awarded, 1864, it was divided between Cremona and Sturm for their answers to a question set by Weierstrass concerning cubic surfaces, a currently active topic. In 1868 the prize was shared between Kortum and H.J.S. Smith for works on cubic and quartic curves in the plane. The competition ran easily enough as long as Steiner's followers were still alive, but soon successors proved hard to find. In 1870 Borchardt proposed the topic of lines of curvature on surfaces, but no essays were forthcoming, and in 1872 the prize went to Hesse for his work in geometry as a whole, and in 1874 to Cremona, again in recognition of his work in general. In 1874 the judges called for entries on the theory of polyhedra, but none were forthcoming and the topic was withdrawn in 1876. Instead, the prize was awarded to H. Schröter for his work extending and deepening Steiner's geometrical methods. The judges then announced a prize of 1,800 marks for an essay on higher algebraic space curves, but there were no entries. The closing date was extended to 1880, and the prize was awarded to Theodor Reye 'for his distinguished work on pure geometry'. In 1880 there was still no satisfactory entry. The judges awarded the prize to L. Lindelöf for a solution to Steiner's problem about the maximum volume of polyhedra of

[7]In 1871 the Thaler was replaced by the German mark, at a rate of 1 Thaler = 3 German marks = 75 U.S. cents. The prize was worth well more than the average annual income of a teacher.

a given type and further extended the closing date for the essay originally set in 1876. Finally, in 1882, they announced two significant essays that were worthy of sharing the prize: Max Noether's and Henri Halphen's. A third essay received an honourable mention, but the author's name was not revealed (most likely it was Rudolf Sturm, who promptly published on that topic). And so it went on. In 1884 and 1888 Fiedler and Zeuthen were rewarded for their distinguished contributions to geometry. Only in 1886 was the prize awarded, to Ernst Kötter, for an essay on the question proposed in 1882 and modified in 1884, which called for a theory of higher curves and surfaces that invoked really existing objects to replace the imaginary points, lines, and planes of contemporary algebraic geometry.

In 1888 Kronecker, with the support of Fuchs, asked that the terms of reference of the prize be changed. This was difficult to achieve, but in late 1889 it was agreed that from 1890 the competition would be announced once every five years, and in the event that no entry was judged satisfactory the prize money could be allocated to significant work, primarily in the field of geometry, written in the previous ten years — a marked relaxation of the original rules. In this spirit, no entries having been received on the set topic (lines of curvature on surfaces, again) Gundelfinger and Schottky shared the prize in 1895. In 1900 Hilbert was awarded a one-third part of the Steiner prize for his work on the foundations of geometry (*Grundlagen der Geometrie*, 1899).

The other winners were Hauck and Lindemann. It was the same story in 1905, when the prize went to Darboux. One is forced to conclude that not even a prize could rescue the methods of synthetic geometry from entering into a prolonged eclipse.

Harnack noted that prize competitions were no longer favoured by the Berlin Academy and declined in importance, and on a number of occasions the prize had to be held back for want of a good enough entry [**24**, p. 397]. In fact, the prize competition organised by the Berlin Academy had terrible results in the 19th century. It got off to an unfortunate start in 1836 when a question set by Dirichlet asking for numerical methods for solving polynomial equations with real or complex roots drew no answers. In 1840 Dirichlet replaced this question with another, inspired by the recent work of Abel, about integrals of algebraic functions. This question also drew no response by the closing date in 1844, and he replaced it with a third, in 1852, where he asked for a proof that the differential equations of dynamics cannot in general be reduced to integrals but require the introduction of new analytic expressions. Dirichlet had his friend Jacobi's work in mind, as he had done in setting the earlier problem, in this case Jacobi's analysis of the spinning top. Yet again there were no answers. In 1858, by which time Dirichlet had moved to Göttingen, Borchardt took up the challenge of

setting an attractive question. He proposed the subject of lines of curvature on surfaces, to no avail: Only one, unsatisfactory, essay was received. Finally, in 1864 Weierstrass proposed the topic of finding a significant problem whose solution requires the new elliptic or Abelian functions, and the academy was able to award the prize for the first time, to Weierstrass's former student Schwarz for his work on minimal surfaces.

Repeated failure must have given the academicians pause, because the prize was not offered again until 1894, when Lazarus Fuchs offered a topic arising out of his own work on differential equations. This attracted no entries, and was re-advertised, in a slightly altered form, in 1894, again without success. In 1902 the prize was awarded to Mertens, a former graduate of Berlin by then in Vienna, for his contributions to mathematics, and Fuchs's question was re-advertised as a question about functions of several variables which are invariant under certain linear transformations. There were no entries, and in 1910 the prize went to Koebe for his work on the uniformisation theorem, and Frobenius asked a question about the class number of the most general cyclotomic field. By 1914 there were no entries, and World War I brought this dismal sequence to an end.

Other societies were more carefully managed. The Jablonowski Society (more properly, the Fürstlich Jablonowshi'schen Gesellschaft der Wissenschaften) was founded in Leipzig by the Polish Prince Jablonowski in 1774 after some years as a private institution. He used it to propose prize problems and sponsored a journal. Problems were set in the domains of mathematics and physics, economics, Polish history, and the history of the Slavs. The society's members were the professors of the University of Leipzig, and they were responsible for running the competitions. In the early years of the 19th century the prizes became unattractive, but the society's finances prospered, and in 1846 it was influential in founding the Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. They benefited from the affair of the Göttingen Seven — seven professors, Gauss's friend and colleague the physicist Wilhelm Weber among them, who were expelled from the university of Göttingen for refusing to accept the terms imposed by the Duke of Cumberland. Weber moved to Leipzig and was one of a number of scientists who built up the university's reputation considerably in the 19th century. The society remained active until 1948, when its leader moved to Jena, and was refounded with help from the Polish government in 1978 with the aim of improving German–Polish economic and cultural relations.

Guided by the Leipzig professors, the Jablonowski Society had more success than many in proposing suitable topics. The first time the prize was awarded after the mid-century reforms, in 1847, it went to H. Grassmann for work connecting geometric analysis to Leibniz's geometric characteristic. In 1884 the society called for essays on the general surface of order 4,

extending the work of Schläfli, Klein, Zeuthen and Rodenburg on cubic surfaces, and gave the prize two years later to Karl Rohn. In 1890 they asked for essays extending the work of Sophus Lie on the invariant theory of arbitrary differential equations, and in 1893 they received an essay from M.A. Tresse that he completed in 1895; he was awarded the prize in 1896. Tresse was a student of Engel's and Lie's in Leipzig. In 1902 they asked for an essay which would essentially complete the work of Poincaré on Neumann's method and the Dirichlet problem. This was a propitious theme, and prizes went to E.R. Neumann in 1905, Plemelj in 1911, Neumann again in 1912, and Gustav Herglotz in 1914.

The Danish Academy of Sciences also awarded prizes from time to time during the 19th century. These seem to have been managed in a traditional way, with titles announced in advance, and in 1823 they had notable success when they awarded Gauss the prize for his essay on the conformal representation of one surface on another. Other prizes were awarded in 1875, to Schubert for his work on the enumerative geometry of cubic curves in space, and to Gram in 1884 for a paper on the prime number theorem. Still, in these cases one notices that the commission charged with conducting the prize had a shrewd eye to success. In the 1820s Schumacher, a prominent geodesist and astronomer, who had organised a survey of Denmark, knew very well that Gauss was conducting a detailed survey of Hanover that was intended to extend the Danish survey, because he was in extensive correspondence with him. What is more, the prize question was formulated by Gauss himself, who then abstained from entering the competition himself for two years before submitting his essay [**7**, p. 102 n.]. In the 1870s the Danish geometer Zeuthen was particularly attracted to enumerative geometry, a subject in which Schubert was emerging as the leading figure. Prizes otherwise seem to have been set almost every year, more often on applied than on pure topics, and generally with little success (only 14 times in the years 1800–1886 for which records are easily accessible).[8]

Jules Henri Poincaré

A rare example of a successful prize competition is the one organised by the Swedish mathematician Mittag-Leffler at the request of the Swedish

[8]I thank Jesper Lützen for this information and his guidance in my reading of the standard source [**29**].

king Oscar II, who wished in this way to mark his 60th birthday, but even this illustrates the perils and pitfalls of such competitions [**3**]. Mittag-Leffler enlisted Weierstrass and Hermite as judges, and together they proposed four topics, while reserving the right to award the prize to any valuable entry on the theory of functions if none of the questions were adequately answered.

King Oscar II of Sweden

Their choice of questions provoked Kronecker to claim that the fourth of these had already been proposed, and indeed that he had solved it, but this allegation eventually petered out. By the closing date twelve entries had been submitted (there was also a late entry from an English angle-trisector). One entry stood out, Henri Poincaré's on the three-body problem. Poincaré had not only submitted a memoir, but he had added to it as time went by in answer to questions about his work that Mittag-Leffler had sent to him. This outraged Weierstrass, who insisted that such irregular behaviour would never have been contemplated in Berlin. In due course Poincaré was declared the winner, with Appell an honourable second.

Poincaré was awarded the prize of 2,500 kroner and a gold medal, and the printers of *Acta Mathematica* were instructed to start printing his revised manuscript.[9] Even at this stage Mittag-Leffler's editorial assistant Phragmén was raising points in the memoir that he did not understand, and in answer to one of these Poincaré admitted that the memoir as it stood was in serious error. Mittag-Leffler ordered the printing halted and all copies of the printed version destroyed, doubtless to prevent his hostile critics on the editorial board of *Acta* and beyond from finding ammunition in the debacle. Then Poincaré found a way to profit from the mistake, and in so doing created the theory of what are now called homoclinic tangles thus opening the way to a mathematical theory of chaos. Mittag-Leffler was happy to print the revised manuscript, but he also charged Poincaré the full printing costs, which came to more than the original prize money. So although the prize competition called forth several good entries and one major paper in the history of mathematics, the stresses involved were too great. No further royal competitions in mathematics were organised, and the king, who had studied mathematics as a young man, found other ways to support mathematicians. For example, he rewarded Kovalevskaya, who

[9]Barrow-Green [**3**, p. 58 n. 101] citing [**16**] observes that Mittag-Leffler's annual salary at the time was 7,000 kroner.

was by then a professor in Stockholm, for a paper extending the work for which she had won the prix Bordin.

Another well-known prize for the solution of a specific problem is the recently awarded Wolfskehl prize, offered for a solution of Fermat's Last Theorem.[10] This prize was first established by Paul Wolfskehl, who came from a wealthy, charitable Jewish banking family. Paul Wolfskehl trained as a doctor but took up the study of mathematics when multiple sclerosis made it clear to him that he would soon be unable to practice. It is very likely that the solace he found in mathematics during the long years of his illness inspired him to create the prize. Wolfskehl died on 13 September 1906, and according to the terms of his will, 100,000 gold marks were set aside for the correct solution of the problem. The Royal Society of Science in Göttingen was charged with administering the fund and adjudicating the solutions. Conditions for the prize were settled and published in 1908, and there was a closing date of almost a century hence: 13 September 2007. A proof of Fermat's Last Theorem, or, if it is false, a characterisation of the exponents for which it is true, would qualify for the prize, but a mere counterexample would not.

From some perspectives, such as generating enthusiasm for mathematics, the prize was a great success; from others, such as the advancement of knowledge, it was a complete disaster. In the first year no fewer than 621 solutions were submitted, and over the years more than 5,000 came in. These had to be read, the errors spotted, and the authors informed, who often replied with attempts to fix their 'proofs'. One can only assume that most, if not all, of the authors knew very much less than Wolfskehl himself about the depth of the problem, but one of them was Ferdinand Lindemann (famous for his proof that π is transcendental) who failed twice, in 1901 and 1908. Much work was also done in Berlin handling correspondence about the prize. Here another doctor with a love for mathematics, Albert Fleck, dealt with so many attempts on behalf of the Berlin Academy of Sciences, Lindemann's among them, that he was eventually awarded the Society's silver Leibniz medal in 1915 for his work; mathematicians in Berlin referred to his operation as the 'Fermat Clinic'. Estimates have varied over the years, too, about the cash value of the prize. It turns out to have been prudently invested and to have survived the strains of high inflation and the vicissitudes of German history better than was often said. Had it been kept safely in gold, its present value would be around $1,400,000. When finally it was awarded to Andrew Wiles in 1997, it was worth DM 75,000 (approximately $37,000) but many estimates had by then written it off and the best put it at around DM 10,000.

[10]This account follows [**2**].

4. The Hilbert Problems

The most successful attempt to reverse the trend toward retrospective prizes and to set problems on topics that would actually bring forth new work is, of course, that of David Hilbert, who proposed 23 problems at the International Congress of Mathematicians in Paris in 1900.[11] His thrilling opening words captured exactly the appeal of great problems: "Who among us," began Hilbert, "would not be glad to lift the veil behind which the future lies hidden; to cast a glance at the next advances of our science and at the secrets of its development during future centuries?" His close friend Minkowski had encouraged him to seize the opportunity to shape the next century in mathematics, writing to him that "Most alluring would be the attempt to look into the future, in other words, a characterisation of the problems to which the mathematicians should turn in the future. With this, you might conceivably have people talking about your speech even decades from now. Of course, prophecy is indeed a difficult thing"[**36**, p. 119].

David Hilbert

The actual speech on the day was something of a disappointment, but in their written, published form the problems gradually worked their charm on the mathematical community. The text is infused with Hilbert's confidence that any problem can be solved. He liked to proclaim on various occasions that there is no 'ignorabimus (we shall not know) in mathematics'. He regarded great problems as crucial for the growth of mathematical knowledge, and he took two as exemplary: Johann Bernoulli's brachistochrone problem, and Fermat's Last Theorem. The first is rooted in empirical sources, the second in the purely mental thought processes of human beings, and creative mathematics, for Hilbert, moves between the two. Thus problem solving and theory formation go hand in hand. The brachistochrone problem had initiated the calculus of variations, a branch of mathematics about to absorb some of Hilbert's own attention. Fermat's Last Theorem had already led to Kummer's work on ideal numbers and thence to the theory of algebraic number fields, the subject of Hilbert's *Zahlbericht* of 1897.

Between these two sources, the applied one augmented by mention of Poincaré's recent solution of the three-body problem, Hilbert placed a num-

[11]See [**23, 41**] and numerous studies of the individual problems.

ber of contemporary developments demonstrating the unity of mathematics as he saw it. Whereas Poincaré on such occasions always sought to emphasise the importance of applications, Hilbert asserted that it was problems rooted in purely mental thought processes that gave rise to practically 'all the finer problems of modern number and function theory.' The result was the miraculous pre-established harmony that the 'mathematician encounters so often in the questions, methods, and ideas of various fields;' Hilbert gave the example of Felix Klein's study of the Platonic solids, which wove a complex theory that connected geometry, group theory, Riemann surfaces, and Galois theory with the theory of linear differential equations.

The specific problems Hilbert raised, and there are more than 23 because several problems come in families, are of various kinds, and cannot all be considered here [**23, 41**]. Some are more like programmes, of which the sixth is the most ambitious. It called for an axiomatisation of physics: Hilbert had recently axiomatised geometry, which he saw as the best-understood branch of science. He imagined that mechanics was ripe for similar treatment and hoped that each branch of science could be dealt with in the same way, because he felt that only an axiomatised theory could respond well to the discoveries made by the experimenters. Indeed, the first six problems form a coherent group focused on foundational questions. The first is the continuum hypothesis, identified by Hilbert as the most interesting problem of the day in set theory. The second calls for foundations for arithmetic, necessary because Hilbert's axiomatisation of geometry rested on otherwise undefended assumptions about number. The first of these was already a significant problem in mathematics, but the second was original with Hilbert and proved difficult to sell, partly because the Italians, who were strong in this area, bridled at Hilbert's intervention, and partly because, at that time, Hilbert had very little idea how it might be solved.

The next six problems belong to number theory and are largely algebraic: the transcendence of certain numbers, the Riemann hypothesis and the distribution of prime numbers, and the solvability of any Diophantine equation may be mentioned here. It is noteworthy that Hilbert's hunches were often wrong. The problem he thought might go first, the Riemann hypothesis, is still with us, but the transcendence questions were solved relatively soon, in the 1930s. The next six are largely geometric and of a more specialist appeal, while the last five are in analysis, the direction in which Hilbert's own interests were going. To cite just three of these, he proposed that a class of what he called 'regular' problems in the calculus of variations should have analytic solutions, he asked for a study of boundary value problems and the Dirichlet problem, and he asked about the general theory of calculus of variations.

The selection of problems was remarkably well done. It is possible, and interesting, to note that entire topics are missing that soon became major areas of 20th century mathematics (topology, measure theory and the Lebesgue integral, for example), but it is more important to note that, in contrast to every learned society in the previous hundred years, Hilbert picked topics that people wanted to work on. Hilbert's problems did not have a closing date of two years hence and a panel of judges appointed to evaluate them; had that been the case there would indeed have been little to show in the first few years. But Hilbert was aiming for longer-term success, and this he achieved. The designated problems are an astute mixture of those known to be important and those deriving from his own experience as a mathematician, which was already broad and was rooted in the fertile soil of the university of Göttingen. Very few seem unduly narrow, and several have been profitably reformulated in light of later experience, a sure sign that Hilbert was on to something deep.

Hilbert shrewdly allowed that a proof that something could not be done counted as a solution, and not as an indication that there are things in mathematics we shall not know. He was also fortunate that many of his problems retained interest even when it became clear that they were not going to have the solution he expected. Solutions, as Hilbert astutely recognised, could also be in the negative, as long as a genuine proof was given that the answer could not be found by the stated means. His paradigm example was the work of Abel and Galois that showed that the quintic equation could not be solved by radicals.

Hilbert spoke of problems in mathematics in terms that can be echoed today. Problems were a sure sign of life. The clarity and ease of comprehension often insisted upon for a mathematical theory, "I should still more demand for a mathematical problem if it is to be perfect; for what is clear and easily comprehended attracts, the complicated repels us. Moreover a mathematical problem should be difficult in order to entice us, yet not completely inaccessible, lest it mock our efforts. It should be to us a signpost on the tortuous paths to hidden truths, ultimately rewarding us by the pleasure in the successful solution." He argued in favour of rigour on the grounds of simplicity, and extended this requirement as far as geometry, mechanics, and even physics. He suggested that both generalisation and specialisation had valuable roles to play in tackling problems.

5. Some Famous Retrospective Prizes

In 1895 the University of Kasan established a prize to recognise the achievements of their distinguished rector in the early days of the university, Nicolai Ivanovich Lobachevskii, one of the discoverers of non-Euclidean geometry. The Lobachevskii prize was to be awarded for the best recent

book on a geometrical subject, and particularly for books on non-Euclidean geometry. It was awarded for the first time in 1897, when it went to Sophus Lie for the third volume of his *Theorie der Transformationsgruppen* [1893]. Klein proposed Lie for the prize, a magnanimous gesture on his part since Lie made what verged on a personal attack on Klein in the preface to that book. The second time the prize was awarded was in 1900, when it went to Wilhelm Killing for the second volume of his *Einführung in der Grundlagen der Geometrie*. In 1904 the prize went to David Hilbert, for whom Poincaré wrote a very strong recommendation adapted from his highly positive review of Hilbert's *Grundlagen der Geometrie*. In this connection it is amusing to note that in 1905 the first award of the Wolfgang Bolyai prize of the Hungarian Academy of Sciences went to Poincaré, while Hilbert received a special citation, and in 1910 the second Bolyai award went to Hilbert with Poincaré again the author of a glowing tribute. Sadly, this prize lapsed during World War I[12] but the Lobachevskii prize continues almost uninterrupted to this day and numbers among its most distinguished recipients Hermann Weyl (1927) and Kolmogorov (1986).

The Nobel prizes were established in the will of Alfred Nobel (1833–1895), in which he created a fund: "the interest on which shall be annually distributed in the form of prizes to those who, during the preceding year, shall have conferred the greatest benefit on mankind." These benefits were to be found in work on physics, chemistry, physiology or medicine, literature, and peace. There is no reason to suppose that Nobel seriously contemplated a Nobel prize in mathematics, which was not and is not self-evidently beneficial to mankind in the way the designated topics are.[13] Nonetheless, Mittag-Leffler does seem to have hoped that Nobel might have donated money to the Swedish Hogskola (the precursor of the University of Stockholm) and to have begun negotiations with him with that aim in mind. He was very disappointed when Nobel did not do so. As Crawford writes: "although it is not known how those in responsible positions at the Hogskola came to believe that a large bequest was forthcoming, this indeed was the expectation, and the disappointment was keen when it was announced early in 1897 that the Hogskola had been left out of Nobel's final will in 1895. Recriminations followed, with both Pettersson and Arrhenius [academic rivals of Mittag-Leffler in the administration of the Hogskola] letting it be known that Nobel's dislike for Mittag-Leffler had brought about what Pettersson termed the 'Nobel Flop' "[**13**, p. 53]. In any case, the Nobel prizes, in line with 19th century experience, were entirely retrospective in nature.

[12]It was revived in 2000, when the prize went to S. Shelah for his *Cardinal Arithmetic*, Oxford University Press, 1994.

[13]Economics was added in 1968, and one may wonder, once the dismal science has been admitted, what else might one day qualify.

During the second half of the 20th century the Fields Medals, awarded every four years at the International Congress of Mathematicians (ICM), established themselves as the most prestigious prize in mathematics. They were established in the will of the Canadian mathematician John Charles Fields (1863–1932). He had been involved in the organisation of the ICM in Toronto in 1924, from which German mathematicians were excluded because passions were still running intensely after World War I. This pained Fields, who had been educated in Germany, so he endowed the medals, which were awarded for the first time at the ICM in Oslo in 1936, four years after his death. The original plan provided for two medals to be awarded at every ICM. The number has since grown on occasion to three or four. Nor was there an explicit statement that the prize be awarded only to people who are under 40, although that has always been the case. These medals are retrospective, as are almost all contemporary prizes.

Bibliography

[1] D.S. Alexander, *A history of complex dynamics, from Schröder to Fatou and Julia*, Vieweg, Aspects of mathematics E. 24, 1994.

[2] K. Barner, *Paul Wolfskehl and the Wolfskehl Prize*, Notices of the AMS **44** (1997), 1294–1303.

[3] J.E. Barrow-Green, *Poincaré and the three body problem*, AMS–LMS, Providence, RI, 1997.

[4] B. Belhoste, *Augustin-Louis Cauchy. A Biography*, translated by F. Ragland, Springer-Verlag, New York, 1991.

[5] K.-R. Biermann, *Aus der geschichte Berliner mathematischer Preisaufgaben*, Wissenschaftliche Zeitschrift der Humboldt-Universität zu Berlin, Mathematische-Naturwissenschaftliche Reihe **13** (1964), 185–197

[6] L.L. Bucciarelli and N. Dworsky, *Sophie Germain. An Essay in the History of the Theory of Elasticity*, D. Reidel, Dordrecht-Boston, MA, 1980.

[7] W.K. Bühler, *Gauss, A Biographical Study*, Springer-Verlag, New York, 1981.

[8] L. Carnot, *Réflections sur la métaphysique du calcul infinitesimal*, Paris, Duprat, 1797.

[9] A.L. Cauchy, *Cours d'analyse de l'École Royale Polytechnique, première partie*, Debure frères, Paris, 1821. Facsimile re-edition, ed. U. Bottazzini, Instrumenta rationis, vol. 7, Clueb, Bologna, 1990

[10] C. Ciliberto, *M. de Francis and the theory of hyperelliptic surfaces*, Supplemento ai rendiconti del circolo matematico di Palermo (2) **55** (1998), 45–74.

[11] R. Cooke, *The Mathematics of Sonya Kovalevskaya*, Springer-Verlag, New York, 1984.

[12] L. Corry, *David Hilbert and the Axiomatization of Physics (1894-1905)*, Archive for History of Exact Sciences **51** (1997), 83–198.

[13] E.R. Crawford, *The Beginnings of the Nobel Institution. The science prizes, 1901-1915*, Cambridge University Press and Editions de la Maison des Sciences de l'Homme, Cambridge, 1984.

[14] E.R. Crawford, *The prize system of the Academy of Sciences, 1850-1914*, Science under control; The French Academy of Sciences, 1795-1914 (M. Crosland, ed.), Cambridge University Press, Cambridge, 1997, 283–307.

[15] J. Dieudonné, *Minkowski, Hermann*, in Dictionary of Scientific Biography, IX, Scribner's Sons, New York, 1990, 140–144.

[16] Y. Domar, *On the foundation of Acta Mathematica*, Acta Math. **148** (1982), 3–8.

[17] H.M. Edwards *Fermat's Last Theorem. A Genetic Introduction to Algebraic Number Theory*, Springer-Verlag, New York, 1997.
[18] J.G. auvel and J.J. Gray, eds., *The History of Mathematics: A Reader*, Macmillan, Basingstoke, Hampshire, 1987.
[19] P. Gauja, *Les fondations de l'Académie des Sciences (1881–1915)*, Hendaye, Basses-Pyrénés, 1915.
[20] J.V. Grabiner, *The Origins of Cauchy's Rigorous Calculus*, M.I.T. Press, Cambridge, MA, 1981.
[21] H.G. Grassmann, *Geometrische Analyse geknüpft an die von Leibniz erfundene geometrische Charakteristik, etc*, Fürstlich Jablonowskische Gesellschaft Preisschriften, Leipzig, 1847.
[22] J.J. Gray, *Linear Differential Equations and Group Theory from Riemann to Poincaré*, Birkhäuser, Boston and Basel, 2000.
[23] J.J. Gray, *The Hilbert Challenge*, Oxford University Press, Oxford, 2000.
[24] A. Harnack, *Geschichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin, im Auftrage der Akademie bearbeitet von Adolf Harnack*, 3 vols, Reichsdruckerei, Berlin, 1900.
[25] D. Hilbert, *Die Theorie der algebraischen Zahlkörper (*Zahlbericht*)*, Jahresbericht den Deutschen mathematiker Vereinigung **4**, 175–546, in *Gesammelte Abhandlungen* **1** (1897), 63–363. English edition, *The Theory of Algebraic Number Fields*, translated and edited by Franz Lemmermeyer and Norbert Schappacher, Springer-Verlag, New York, 1998.
[26] J. Itard, *Lagrange, Joseph Louis*, in Dictionary of Scientific Biography, VII, Scribner's Sons, New York, 1975, 559–573.
[27] S. L'Huillier, *Exposition élémentaire des principes des calculs supérieure*, Decker, Berlin, 1787.
[28] J.L. Lagrange, *Théorie des fonctions analytiques*, Impr. de la République, Paris, 1797.
[29] A. Lomholt, *Det Kongelige Danske Videnskabers Selskab, 1749–1942*, I kommission hos E. Munksgaard, Copenhagen, 1942.
[30] J. Lützen, *Joseph Liouville 1809–1882, Master of Pure and Applied Mathematics*, Springer-Verlag, New York, 1990.
[31] C. MacLaurin, *A Treatise of Fluxions in two Books*, Ruddimans, Edinburgh, 1742.
[32] E. Maindron, *Les fondations de prix à l'Académie des sciences. Les laureates de l'Académie, 1714-1880*, Gauthier-Villars, Paris, 1881.
[33] V. Maz'ya and T. Shaposhnikova, *Jacques Hadamard, A Universal Mathematician*, AMS-LMS, Providence, RI, 1998.
[34] M. Monastyrsky, *Modern Mathematics in the Light of the Fields Medals*, A.K. Peters, Wellesley, Mass, 1996.
[35] H. Poincaré, *Three Supplementary Essays on the Discovery of Fuchsian Functions* (J.J. Gray and S. Walter, eds.), Akademie Verlag, Berlin and Blanchard, Paris, 1997.
[36] L. Rüdenberg and H. Zassenhaus (eds.), *Hermann Minkowski — Briefe an David Hilbert*, Springer-Verlag, Berlin, 1973.
[37] M. du Sautoy, *The Music of the Primes. Searching to Solve the Greatest Mystery in Mathematics* HarperCollins, New York, 2003.
[38] S. Smale, *Mathematical problems for the next century*, in Mathematics: Frontiers and Perspectives (V. Arnold, M. Atiyah, P. Lax, and B. azur, eds), AMS, Providence, RI, 2000.
[39] R. Taton, *Euler et d'Alembert*, in Zum Werk Leonhard Eulers (E. Knobloch et al, eds.) Birkhäuser Verlag, Boston and Basel, 1984, 95–117.
[40] R.S. Westfall, *Never at Rest*, Cambridge University Press, Cambridge, 1980.
[41] B.H. Yandell, *The Honors Class. Hilbert's Problems and Their Solvers*, A.K. Peters, Natick, MA, 2001.

The Birch and Swinnerton-Dyer Conjecture

Andrew Wiles

The Birch and Swinnerton-Dyer Conjecture

Andrew Wiles

A polynomial relation $f(x, y) = 0$ in two variables defines a curve C_0. If the coefficients of the polynomial are rational numbers, then one can ask for solutions of the equation $f(x, y) = 0$ with $x, y \in \mathbb{Q}$, in other words for rational points on the curve. If we consider a non-singular projective model C of the curve, then topologically it is classified by its genus, and we call this the genus of C_0 also. Note that $C_0(\mathbb{Q})$ and $C(\mathbb{Q})$ are either both finite or both infinite. Mordell conjectured, and in 1983 Faltings proved, the following deep result.

Theorem ([**9**]). *If the genus of C_0 is greater than or equal to 2, then $C_0(\mathbb{Q})$ is finite.*

As yet the proof is not effective so that one does not possess an algorithm for finding the rational points. (There is an effective bound on the number of solutions but that does not help much with finding them.)

The case of genus zero curves is much easier and was treated in detail by Hilbert and Hurwitz [**12**]. They explicitly reduce to the cases of linear and quadratic equations. The former case is easy and the latter is resolved by the criterion of Legendre. In particular, for a non-singular projective model C we find that $C(\mathbb{Q})$ is non-empty if and only if C has p-adic points for all primes p, and this in turn is determined by a finite number of congruences. If $C(\mathbb{Q})$ is non-empty, then C is parametrized by rational functions and there are infinitely many rational points.

The most elusive case is that of genus 1. There may or may not be rational solutions and no method is known for determining which is the case for any given curve. Moreover when there are rational solutions there may or may not be infinitely many. If a non-singular projective model C has a rational point, then $C(\mathbb{Q})$ has a natural structure as an abelian group with this point as the identity element. In this case we call C an elliptic curve over $\mathbb{Q}$. (For a history of the development of this idea see [**19**].) In 1922 Mordell [**15**] proved that this group is finitely generated, thus fulfilling an implicit assumption of Poincaré.

THEOREM. *If C is an elliptic curve over $\mathbb{Q}$, then*

$$C(\mathbb{Q}) \simeq \mathbb{Z}^r \oplus C(\mathbb{Q})^{\text{tors}}$$

for some integer $r \geq 0$, where $C(\mathbb{Q})^{\text{tors}}$ is a finite abelian group.

The integer r is called the rank of C. It is zero if and only if $C(\mathbb{Q})$ is finite. We can find an affine model for the curve in Weierstrass form

$$C\colon y^2 = x^3 + ax + b$$

with $a, b \in \mathbb{Z}$. We let Δ denote the discriminant of the cubic and set

$$\begin{aligned} N_p &:= \#\{\text{solutions of } y^2 \equiv x^3 + ax + b \mod p\}, \\ a_p &:= p - N_p. \end{aligned}$$

Then we can define the incomplete L-series of C (incomplete because we omit the Euler factors for primes $p|2\Delta$) by

$$L(C, s) := \prod_{p \nmid 2\Delta} (1 - a_p p^{-s} + p^{1-2s})^{-1}.$$

We view this as a function of the complex variable s and this Euler product is then known to converge for $\text{Re}(s) > 3/2$. A conjecture going back to Hasse (see the commentary on 1952(d) in [**26**]) predicted that $L(C, s)$ should have a holomorphic continuation as a function of s to the whole complex plane. This has now been proved ([**25**], [**24**], [**1**]). We can now state the millenium prize problem:

CONJECTURE (Birch and Swinnerton-Dyer). *The Taylor expansion of $L(C, s)$ at $s = 1$ has the form*

$$L(C, s) = c(s - 1)^r + \textit{higher order terms}$$

with $c \neq 0$ and $r = \text{rank}(C(\mathbb{Q}))$.

In particular this conjecture asserts that $L(C, 1) = 0 \Leftrightarrow C(\mathbb{Q})$ is infinite.

REMARKS. 1. There is a refined version of this conjecture. In this version one has to define Euler factors at primes $p|2\Delta$ to obtain the completed L-series, $L^*(C, s)$. The conjecture then predicts that $L^*(C, s) \sim c^*(s - 1)^r$ with

$$c^* = |\text{Ш}_C| R_\infty w_\infty \prod_{p|2\Delta} w_p / |C(\mathbb{Q})^{\text{tors}}|^2.$$

Here $|\text{Ш}_C|$ is the order of the Tate–Shafarevich group of the elliptic curve C, a group which is not known in general to be finite although it is conjectured to be so. It counts the number of equivalence classes of homogeneous spaces of C which have points in all local fields. The term R_∞ is an $r \times r$ determinant whose matrix entries are given by a height pairing applied to a system of generators of $C(\mathbb{Q})/C(\mathbb{Q})^{\text{tors}}$. The w_p's are elementary local factors and w_∞ is a simple multiple of the real period of C. For a precise definition of these

factors see [**20**] or [**22**]. It is hoped that a proof of the conjecture would also yield a proof of the finiteness of $Ш_C$.

2. The conjecture can also be stated over any number field as well as for abelian varieties, see [**20**]. Since the original conjecture was stated, much more elaborate conjectures concerning special values of L-functions have appeared, due to Tate, Lichtenbaum, Deligne, Bloch, Beilinson and others, see [**21**], [**3**] and [**2**]. In particular, these relate the ranks of groups of algebraic cycles to the order of vanishing (or the order of poles) of suitable L-functions.

3. There is an analogous conjecture for elliptic curves over function fields. It has been proved in this case by Artin and Tate [**20**] that the L-series has a zero of order at least r, but the conjecture itself remains unproved. In the function field case it is now known to be equivalent to the finiteness of the Tate–Shafarevich group, [**20**], [**17**, Corollary 9.7].

4. A proof of the conjecture in the stronger form would give an effective means of finding generators for the group of rational points. Actually, one only needs the integrality of the term $Ш_C$ in the expression for $L^*(C, s)$ above, without any interpretation as the order of the Tate–Shafarevich group. This was shown by Manin [**16**] subject to the condition that the elliptic curves were modular, a property which is now known for all elliptic curves by [**25**], [**24**], [**1**]. (A modular elliptic curve is one that occurs as a factor of the Jacobian of a modular curve.)

1. Early History

Problems on curves of genus 1 feature prominently in Diophantus' *Arithmetica*. It is easy to see that a straight line meets an elliptic curve in three points (counting multiplicity) so that if two of the points are rational then so is the third.[1] In particular, if a tangent is taken at a rational point, then it meets the curve again in a rational point. Diophantus implicitly used this method to obtain a second solution from a first. He did not iterate this process, however, and it was Fermat who first realized that one can sometimes obtain infinitely many solutions in this way. Fermat also introduced a method of 'descent' that sometimes permits one to show that the number of solutions is finite or even zero.

One very old problem concerned with rational points on elliptic curves is the congruent number problem. One way of stating it is to ask which rational integers can occur as the areas of right-angled triangles with rational length sides. Such integers are called congruent numbers. For example, Fibonacci was challenged in the court of Frederic II with the problem for $n = 5$,

[1]This was apparently first explicitly pointed out by Newton.

and he succeeded in finding such a triangle. He claimed, moreover, that there was no such triangle for $n = 1$, but the proof was fallacious and the first correct proof was given by Fermat. The problem dates back to Arab manuscripts of the 10th century (for the history see [**27**, Chapter 1, §VII] and [**7**, Chapter XVI]). It is closely related to the problem of determining the rational points on the curve $C_n\colon y^2 = x^3 - n^2x$. Indeed,

$$C_n(\mathbb{Q}) \text{ is infinite } \iff n \text{ is a congruent number.}$$

Assuming the Birch and Swinnerton-Dyer conjecture (or even the weaker statement that $C_n(\mathbb{Q})$ is infinite $\Leftrightarrow L(C_n, 1) = 0$) one can show that any $n \equiv 5, 6, 7 \mod 8$ is a congruent number, and, moreover, Tunnell has shown, again assuming the conjecture, that for n odd and square-free

$$n \text{ is a congruent number } \iff \begin{array}{l} \#\{x, y, z \in \mathbb{Z}\colon 2x^2 + y^2 + 8z^2 = n\} \\ = 2 \times \#\{x, y, z \in \mathbb{Z}\colon 2x^2 + y^2 + 32z^2 = n\}, \end{array}$$

with a similar criterion if n is even [**23**]. Tunnell proved the implication left to right unconditionally with the help of the main theorem of [**5**] described below.

2. Recent History

It was the 1901 paper of Poincaré that started the modern theory of rational points on curves and that first raised questions about the minimal number of generators of $C(\mathbb{Q})$. The conjecture itself was first stated in the form we have given in the early 1960s (see [**4**]). It was found experimentally using one of the early EDSAC computers at Cambridge. The first general result proved was for elliptic curves with complex multiplication. The curves with complex multiplication fall into a finite number of families including $\{y^2 = x^3 - Dx\}$ and $\{y^2 = x^3 - k\}$ for varying $D, k \neq 0$. This theorem was proved in 1976 and is due to Coates and Wiles [**5**]. It states that if C is a curve with complex multiplication and $L(C, 1) \neq 0$, then $C(\mathbb{Q})$ is finite. In 1983 Gross and Zagier showed that if C is a modular elliptic curve and $L(C, 1) = 0$ but $L'(C, 1) \neq 0$, then an earlier construction of Heegner actually gives a rational point of infinite order. Using new ideas together with this result, Kolyvagin showed in 1990 that for modular elliptic curves, if $L(C, 1) \neq 0$ then $r = 0$ and if $L(C, 1) = 0$ but $L'(C, 1) \neq 0$ then $r = 1$. In the former case Kolyvagin needed an analytic hypothesis which was confirmed soon afterwards; see [**6**] for the history of this and for further references. Finally as noted in remark 4 above it is now known that all elliptic curves over $\mathbb{Q}$ are modular so that we now have the following result:

THEOREM. *If $L(C, s) \sim c(s - 1)^m$ with $c \neq 0$ and $m = 0$ or 1, then the conjecture holds.*

In the cases where $m = 0$ or 1 some more precise results on c (which of course depends on the curve) are known by work of Rubin and Kolyvagin.

3. Rational Points on Higher-Dimensional Varieties

We began by discussing the diophantine properties of curves, and we have seen that the problem of giving a criterion for whether $C(\mathbb{Q})$ is finite or not is only an issue for curves of genus 1. Moreover, according to the conjecture above, in the case of genus 1, $C(\mathbb{Q})$ is finite if and only if $L(C, 1) \neq 0$. In higher dimensions, if V is an algebraic variety, it is conjectured (see [**14**]) that if we remove from V (the closure of) all subvarieties that are images of $\mathbb{P}^1$ or of abelian varieties, then the remaining open variety W should have the property that $W(\mathbb{Q})$ is finite. This has been proved by Faltings in the case where V is itself a subvariety of an abelian variety [**10**].

This suggests that to find infinitely many points on V one should look for rational curves or abelian varieties in V. In the latter case we can hope to use methods related to the Birch and Swinnerton-Dyer conjecture to find rational points on the abelian variety. As an example of this, consider the conjecture of Euler from 1769 that $x^4 + y^4 + z^4 = t^4$ has no non-trivial solutions. By finding a curve of genus 1 on the surface and a point of infinite order on this curve, Elkies [**8**] found the solution

$$2682440^4 + 15365639^4 + 18796760^4 = 20615673^4.$$

His argument shows that there are infinitely many solutions to Euler's equation.

In conclusion, although there has been some success in the last fifty years in limiting the number of rational points on varieties, there are still almost no methods for finding such points. It is to be hoped that a proof of the Birch and Swinnerton-Dyer conjecture will give some insight concerning this general problem.

Bibliography

[1] C. Breuil, B. Conrad, F. Diamond, and R. Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises*, J. Amer. Math. Soc. **14** (2001), 843–939.

[2] A. Beilinson, *Notes on absolute Hodge cohomology* in Applications of algebraic K-theory to algebraic geometry and number theory, Contemp. Math. **55**, AMS, Providence, RI, 1986, 35–68.

[3] S. Bloch, *Height pairings for algebraic cycles*, J. Pure Appl. Algebra **34** (1984), 119–145.

[4] B. Birch and H. Swinnerton-Dyer, *Notes on elliptic curves* II, Journ. reine u. angewandte Math. **218** (1965), 79–108.

[5] J. Coates and A. Wiles, *On the conjecture of Birch and Swinnerton-Dyer*, Invent. Math. **39** (1977), 223–251.

[6] H. Darmon, *Wiles' theorem and the arithmetic of elliptic curves*, in Modular forms and Fermat's Last Theorem, Springer, Berlin, 1997, 549–569

[7] L. Dickson, *History of the Theory of Numbers*, vol. II, Carnegie Institute, Washington, 1919, reprinted AMS, Providence, RI, 1999.
[8] N. Elkies, *On* $A^4 + B^4 + C^4 = D^4$, Math. Comput. **51** (1988), 825–835.
[9] G. Faltings, *Endlichkeitsätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), 549–576.
[10] G. Faltings, *The general case of S. Lang's conjecture*, in Barsotti Symposium in Algebraic Geometry, Perspec. Math., vol. 15, Academic Press, Boston, 1994, 175–182.
[11] B. Gross and D. Zagier, *Heegner points and derivatives of L-series*, Invent. Math. **84** (1986), 225–320.
[12] D. Hilbert, A. Hurwitz, *Uber die diophantischen Gleichungen von Geschlect Null*, Acta Mathematica **14** (1890), 217–224.
[13] V. Kolyvagin, *Finiteness of* $E(\mathbb{Q})$ *and* $Ш(E, \mathbb{Q})$ *for a class of Weil curves*, Math. USSR, Izv. **32** (1989), 523–541.
[14] S. Lang, *Number Theory III*, Encyclopædia of Mathematical Sciences, vol. 60, Springer-Verlag, Heidelberg, 1991.
[15] L. Mordell, *On the rational solutions of the indeterminate equations of the third and fourth degrees*, Proc. Cambridge Phil. Soc. **21** (1922-23), 179–192.
[16] Y. Manin, *Cyclotomic fields and modular curves*, Russian Mathematical Surveys **26**, no. 6 (1971), 7–78.
[17] J. Milne, *Arithmetic Duality Theorems*, Academic Press, Boston, 1986.
[18] H. Poincaré, *Sur les propriétés arithmétiques des courbes algébriques*, Jour. Math. Pures Appl. **7**, Ser. 5 (1901).
[19] N. Schappacher, *Développement de la loi de groupe sur une cubique*, in Seminaire de Théorie des Nombres, Paris 1988/89, Progress in Math., vol. 91, Birkhaüser, Basel, 1991, 159–184.
[20] J. Tate, *On the conjectures of Birch and Swinnerton-Dyer and a geometric analog*, Seminaire Bourbaki 1965/66, no. 306.
[21] J. Tate, *Algebraic cycles and poles of zeta functions*, in Arithmetical Algebraic Geometry: Proceedings of a Conference at Purdue University, O.F.G. Schilling, ed., Harper and Row, New York, 1965, 93–110.
[22] J. Tate, *The arithmetic of elliptic curves*, Invent. Math. **23** (1974), 179–206.
[23] J. Tunnell, *A classical diophantine problem and modular forms of weight* 3/2, Invent. Math. **72** (1983), 323–334.
[24] R. Taylor, A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. Math. **141** (1995), 553–572.
[25] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. Math. **142** (1995), 443–551.
[26] A. Weil, *Collected Papers*, Vol. II, Springer-Verlag, New York, 1979.
[27] A. Weil, *Basic Number Theory*, Birkhaüser, Boston, 1984.

The Hodge Conjecture

Pierre Deligne

The Hodge Conjecture

Pierre Deligne

1. Statement

We recall that a pseudo complex structure on a C^∞-manifold X of dimension $2N$ is a $\mathbb{C}$-module structure on the tangent bundle T_X. Such a module structure induces an action of the group $\mathbb{C}^*$ on T_X, with $\lambda \in \mathbb{C}^*$ acting by multiplication by λ. By transport of structures, the group $\mathbb{C}^*$ acts also on each exterior power $\wedge^n T_X$, as well as on the complexified dual $\Omega^n := \mathcal{H}\mathrm{om}(\wedge^n T_X, \mathbb{C})$. For $p+q=n$, a (p,q)-*form* is a section of Ω^n on which $\lambda \in \mathbb{C}^*$ acts by multiplication by $\lambda^{-p}\bar{\lambda}^{-q}$.

From now on, we assume X complex analytic. A (p,q)-form is then a form which, in local holomorphic coordinates, can be written as

$$\sum a_{i_1,\dots,i_p,j_1\dots j_q} dz_{i_1}\wedge\cdots\wedge dz_{i_p}\wedge d\bar{z}_{j_1}\wedge\cdots\wedge d\bar{z}_{j_q},$$

and the decomposition $\Omega^n = \oplus\Omega^{p,q}$ induces a decomposition $d = d' + d''$ of the exterior differential, with d' (resp. d'') of degree $(1,0)$ (resp. $(0,1)$).

If X is compact and admits a Kähler metric, for instance if X is a projective non-singular algebraic variety, this action of $\mathbb{C}^*$ on forms induces an action on cohomology. More precisely, $H^n(X,\mathbb{C})$ is the space of closed n-forms modulo exact forms, and if we define $H^{p,q}$ to be the space of closed (p,q)-forms modulo the $d'd''$ of $(p-1,q-1)$-forms, the natural map

$$\bigoplus_{p+q=n} H^{p,q} \to H^n(X,\mathbb{C}) \tag{1}$$

is an isomorphism. If we choose a Kähler structure on X, one can give the following interpretation to the decomposition (1) of $H^n(X,\mathbb{C})$: the action of $\mathbb{C}^*$ on forms commutes with the Laplace operator, hence induces an action of $\mathbb{C}^*$ on the space $\mathcal{H}^n$ of harmonic n-forms. We have $\mathcal{H}^n \xrightarrow{\sim} H^n(X,\mathbb{C})$ and $H^{p,q}$ identifies with the space of harmonic (p,q)-forms.

When X moves in a holomorphic family, the *Hodge filtration* $F^p := \bigoplus_{a\geq p} H^{a,n-a}$ of $H^n(X,\mathbb{C})$ is better behaved than the Hodge decomposition. Locally on the parameter space T, $H^n(X_t,\mathbb{C})$ is independent of $t\in T$ and the Hodge filtration can be viewed as a variable filtration $F(t)$ on a fixed vector

space. It varies holomorphically with t, and obeys Griffiths transversality: at first order around $t_0 \in T$, $F^p(t)$ remains in $F^{p-1}(t_0)$.

So far, we have computed cohomology using C^∞ forms. We could as well have used forms with generalized functions coefficients, that is, currents. The resulting groups $H^n(X, \mathbb{C})$ and $H^{p,q}$ are the same. If Z is a closed analytic subspace of X, of complex codimension p, Z is an integral cycle and, by Poincaré duality, defines a class $\mathrm{cl}(Z)$ in $H^{2p}(X, \mathbb{Z})$. The integration current on Z is a closed (p,p)-form with generalized function coefficients, representing the image of $\mathrm{cl}(Z)$ in $H^{2p}(X, \mathbb{C})$. The class $\mathrm{cl}(Z)$ in $H^{2p}(X, \mathbb{Z})$ is hence of type (p,p), in the sense that its image in $H^{2p}(X, \mathbb{C})$ is. Rational (p,p)-classes are called *Hodge classes*. They form the group

$$H^{2p}(X, \mathbb{Q}) \cap H^{p,p}(X) = H^{2p}(X, \mathbb{Q}) \cap F^p \subset H^{2p}(X, \mathbb{C}).$$

In [**6**], Hodge posed the

Hodge Conjecture. *On a projective non-singular algebraic variety over* $\mathbb{C}$*, any Hodge class is a rational linear combination of classes* $\mathrm{cl}(Z)$ *of algebraic cycles.*

2. Remarks

(i) By Chow's theorem, on a complex projective variety, algebraic cycles are the same as closed analytic subspaces.

(ii) On a projective non-singular variety X over $\mathbb{C}$, the group of integral linear combinations of classes $\mathrm{cl}(Z)$ of algebraic cycles coincides with the group of integral linear combinations of products of Chern classes of algebraic (equivalently by GAGA: analytic) vector bundles. To express $\mathrm{cl}(Z)$ in terms of Chern classes, one resolves the structural sheaf $\mathcal{O}_Z$ by a finite complex of vector bundles. That Chern classes are algebraic cycles holds, basically, because vector bundles have plenty of meromorphic sections.

(iii) A particular case of (ii) is that the integral linear combinations of classes of divisors (= codimension 1 cycles) are simply the first Chern classes of line bundles. If $Z^+ - Z^-$ is the divisor of a meromorphic section of $\mathcal{L}$, $c_1(\mathcal{L}) = \mathrm{cl}(Z^+) - \mathrm{cl}(Z^-)$. This is the starting point of the proof given by Kodaira and Spencer [**7**] of the Hodge conjecture for H^2: a class $c \in H^2(X, \mathbb{Z})$ of type $(1,1)$ has image 0 in the quotient $H^{0,2} = H^2(X, \mathcal{O})$ of $H^2(X, \mathbb{C})$, and the long exact sequence of cohomology defined by the exponential exact sequence

$$0 \longrightarrow \mathbb{Z} \longrightarrow \mathcal{O} \xrightarrow{\exp(2\pi i \quad)} \mathcal{O}^* \longrightarrow 0$$

shows that c is the first Chern class of a line bundle.

(iv) The relation between algebraic cycles and algebraic vector bundles is also the basis of the Atiyah and Hirzebruch theorem [**2**] that the Hodge conjecture cannot hold integrally. In the Atiyah–Hirzebruch spectral sequence from cohomology to topological K-theory,

$$E_2^{pq} = H^p(X, K^q(P^{\underline{t}})) \Longrightarrow K^{p+q}(X);$$

the resulting filtration of $K^n(X)$ is by the

$$F^pK^n(X) = \mathrm{Ker}(K^n(X) \to K^n((p-1)\text{-skeleton, in any triangulation})).$$

Equivalently, a class c is in F^p if for some topological subspace Y of codimension p, c is the image of a class $\tilde{c}$ with support in Y. If Z is an algebraic cycle of codimension p, a resolution of $\mathcal{O}_Z$ defines a K-theory class with support in Z: $c_Z \in K^0(X, X - Z)$. Its image in $F^pK^0(X)$ agrees with the class of Z in $H^{2p}(X, \mathbb{Z})$. The latter hence is in the kernel of the successive differentials d_r of the spectral sequence.

No counterexample is known to the statement that integral (p, p) classes killed by all d_r are integral linear combinations of classes $\mathrm{cl}(Z)$. One has no idea of which classes should be effective, that is, of the form $\mathrm{cl}(Z)$, rather than a difference of such.

On a Stein manifold X, any topological complex vector bundle can be given a holomorphic structure and, at least for X of the homotopy type of a finite CW complex, it follows that any class in $H^{2p}(X, \mathbb{Z})$ in the kernel of all d_r is a $\mathbb{Z}$-linear combination of classes of analytic cycles.

(v) The assumption in the Hodge conjecture that X be algebraic cannot be weakened to X being merely Kähler. See Zucker's appendix to [**11**] for counterexamples where X is a complex torus.

(vi) When Hodge formulated his conjecture, he had not realized it could hold only rationally (i.e. after tensoring with $\mathbb{Q}$). He also proposed a further conjecture, characterizing the subspace of $H^n(X, \mathbb{Z})$ spanned by the images of cohomology classes with support in a suitable closed analytic subspace of complex codimension k. Grothendieck observed that this further conjecture is trivially false, and gave a corrected version of it in [**5**].

3. The Intermediate Jacobian

The cohomology class of an algebraic cycle Z of codimension p has a natural lift to a group $J_p(X)$, extension of the group of classes of type (p, p) in $H^{2p}(X, \mathbb{Z})$ by the *intermediate jacobian*

$$J_p(X)^0 := H^{2p-1}(X, \mathbb{Z}) \setminus H^{2p-1}(X, \mathbb{C})/F^p.$$

This expresses that the class can be given an integral description (in singular cohomology), as well as an analytic one, as a closed (p, p) current, giving a

hypercohomology class in $\mathbb{H}^{2p}$ of the subcomplex $F^p\Omega^*_{\text{hol}} := (0 \to \cdots \to 0 \to \Omega^p_{\text{hol}} \to \cdots)$ of the holomorphic de Rham complex, with an understanding at the cocycle level of why the two agree in $H^{2p}(X, \mathbb{C})$. 'Understanding' means a cochain in a complex computing $H^*(X, \mathbb{C})$, whose coboundary is the difference between cocycles coming from the integral, resp. analytic, constructions. Indeed, $J_p(X)$ is the hypercohomology $\mathbb{H}^{2p}$ of the homotopy kernel of the difference map $\mathbb{Z} \oplus F^p\Omega^*_{\text{hol}} \to \Omega^*$.

In general, using that all algebraic cycles on X fit in a denumerable number of algebraic families, one checks that the subgroup $A_p(X)$ of $J_p(X)$ generated by algebraic cycles is the extension of a denumerable group by its connected component $A^0_p(X)$, and that for some sub-Hodge structure H_{alg} of type $\{(p-1,p),(p,p-1)\}$ of $H^{2p-1}(X)$, $A^0_p(X)$ is $H_{\text{alg}_{\mathbb{Z}}} \setminus H_{\text{alg}_{\mathbb{C}}}/F^p$. 'Sub-Hodge structure' means the subgroup of the integral lattice whose complexification is the sum of its intersections with the $H^{a,b}$. The Hodge conjecture (applied to the product of X and a suitable abelian variety) predicts that H_{alg} is the largest sub-Hodge structure of $H^{2p-1}(X)$ of type $\{(p-1,p)(p,p-1)\}$.

No conjecture is available to predict what subgroup of $J_p(X)$ the group $A_p(X)$ is. Cases are known where $A_p(X)/A^0_p(X)$ is of infinite rank. See, for instance, the paper [**9**] and the references it contains. This has made generally inapplicable the methods introduced by Griffiths (see, for instance, Zucker [**11**]) to prove the Hodge conjecture by induction on the dimension of X, using a Lefschetz pencil of hyperplane sections of X. Indeed, the method requires not just the Hodge conjecture for the hyperplane sections H, but that all of $J_p(H)$ comes from algebraic cycles.

4. Detecting Hodge Classes

Let $(X_s)_{s\in S}$ be an algebraic family of projective non-singular algebraic varieties: the fibers of a projective and smooth map $f\colon X \to S$. We assume it is defined over the algebraic closure $\bar{\mathbb{Q}}$ of $\mathbb{Q}$ in $\mathbb{C}$. No algorithm is known to decide whether a given integral cohomology class of a typical fiber X_0 is somewhere on S of type (p,p). The Hodge conjecture implies that the locus where this happens is a denumerable union of algebraic subvarieties of S (known: see [**4**]), and is defined over $\bar{\mathbb{Q}}$ (unknown).

The Hodge conjecture is not known even in the following nice examples.

Example 1. For X of complex dimension N, the diagonal Δ of $X \times X$ is an algebraic cycle of codimension N. The Hodge decomposition being compatible with Künneth, the Künneth components $\text{cl}(\Delta)_{a,b} \in H^a(X) \otimes H^b(X) \subset H^{2N}(X \times X)$ $(a+b=2N)$ of $\text{cl}(\Delta)$ are Hodge classes.

Example 2. If $\eta \in H^2(X, \mathbb{Z})$ is the cohomology class of a hyperplane section of X, the iterated cup product $\eta^p\colon H^{N-p}(X, \mathbb{C}) \to H^{N+p}(X, \mathbb{C})$ is

an isomorphism (hard Lefschetz theorem, proved by Hodge. See [**10**, IV.6]). Let $\mathfrak{z} \in H^{N-p}(X,\mathbb{C}) \otimes H^{N-p}(X,\mathbb{C}) \subset H^{2N-2p}(X \times X)$ be the class such that the inverse isomorphism $(\eta^p)^{-1}$ is $c \mapsto \mathrm{pr}_{1!}(\mathfrak{z} \cup \mathrm{pr}_2^* c)$. The class $\mathfrak{z}$ is Hodge.

5. Motives

Algebraic varieties admit a panoply of cohomology theories, related over $\mathbb{C}$ by comparison isomorphisms. Resulting structures on $H^*(X,\mathbb{Z})$ should be viewed as analogous to the Hodge structure. Examples: If X is defined over a subfield K of $\mathbb{C}$, with algebraic closure $\bar{K}$ in $\mathbb{C}$, $\mathrm{Gal}(\bar{K}/K)$ acts on $H^*(X,\mathbb{Z}) \otimes \mathbb{Z}_\ell$ and $H^*(X,\mathbb{C}) = H^*(X,\mathbb{Z}) \otimes \mathbb{C}$ has a natural K-structure $H_{\mathrm{DR}}(X \text{ over } K)$, compatible with the Hodge filtration. Those cohomology theories give rise to conjectures parallel to the Hodge conjecture, determining the linear span of classes of algebraic cycles. Example: the Tate conjecture [**8**]. Those conjectures are open even for H^2.

Grothendieck's theory of motives aims at understanding the parallelism between those cohomology theories. Progress is blocked by a lack of methods to construct interesting algebraic cycles. If the cycles of Examples 1 and 2 of §4 were algebraic, Grothendieck's motives over $\mathbb{C}$ would form a semi-simple abelian category with a tensor product, and be the category of representations of some pro-reductive group-scheme. If the algebraicity of those cycles is assumed, the full Hodge conjecture is equivalent to a natural functor from the category of motives to the category of Hodge structures being fully faithful.

6. Substitutes and Weakened Forms

In despair, efforts have been made to find substitutes for the Hodge conjecture. On abelian varieties, Hodge classes at least share many properties of cohomology classes of algebraic cycles: they are "absolutely Hodge" [**3**], even "motivated" [**1**]. This suffices for some applications — for instance, the proof of algebraic relations among periods and quasi periods of abelian varieties predicted by the Hodge conjecture [**3**], but does not allow for reduction modulo p. The following corollaries of the Hodge conjecture would be particularly interesting. Let A be an abelian variety over the algebraic closure $\mathbb{F}$ of a finite field $\mathbb{F}_q$. Lift A in two different ways to characteristic 0, to complex abelian varieties A_1 and A_2 defined over $\bar{\mathbb{Q}}$. Pick Hodge classes $\mathfrak{z}_1$ and $\mathfrak{z}_2$ on A_1 and A_2, of complementary dimension. Interpreting $\mathfrak{z}_1$ and $\mathfrak{z}_2$ as ℓ-adic cohomology classes, one can define the intersection number κ of the reduction of $\mathfrak{z}_1$ and $\mathfrak{z}_2$ over $\mathbb{F}$. Is κ a rational number? If $\mathfrak{z}_1$ and $\mathfrak{z}_2$ were $\mathrm{cl}(Z_1)$ and $\mathrm{cl}(Z_2)$, Z_1 and Z_2 could be chosen to be defined over $\bar{\mathbb{Q}}$ and κ would be the intersection number of the reductions of Z_1 and Z_2. Same

question for the intersection number of the reduction of $\mathfrak{z}_1$ over $\mathbb{F}$ with the class of an algebraic cycle on A.

Bibliography

[1] Y. André, *Pour une théorie inconditionnelle des motifs*, Publ. Math. IHES **83** (1996), 5–49.

[2] M.F. Atiyah and F. Hirzebruch, *Analytic cycles on complex manifolds*, Topology **1** (1962), 25–45.

[3] P. Deligne (rédigé par J. L. Brylinski), *Cycles de Hodge absolus et périodes des intégrales des variétés abéliennes*, Mémoires SMF **2** (1980), 23–33.

[4] P. Deligne, E. Cattani, and A. Kaplan, *On the locus of Hodge classes*, JAMS **8** (1995), 483–505.

[5] A. Grothendieck, *Hodge's general conjecture is false for trivial reasons*, Topology **8** (1969), 299–303.

[6] W.V.D. Hodge, *The topological invariants of algebraic varieties*, in Proceedings ICM 1950, AMS, Providence, RI, 1952, 181–192.

[7] K. Kodaira and D.C. Spencer, *Divisor classes on algebraic varieties*, Proc. Nat. Acad. Sci. **39** (1953), 872–877.

[8] J. Tate, *Algebraic cycles and poles of zeta functions*, in Arithmetic Algebraic Geometry, Harper and Row, New York, 1965, 93–110.

[9] C. Voisin, *The Griffiths group of a general Calabi–Yau threefold is not finitely generated*, Duke Math. J. **102**, 151–186.

[10] A. Weil, *Introduction à l'étude des variétés kahlériennes*, Publ. Univ. Nancago VI, Act. Sci. et Ind. **1267**, Hermann, Paris, 1958.

[11] S. Zucker, *The Hodge conjecture for cubic fourfolds*, Comp. Math. **34** (1977), 199–209.

Existence and Smoothness of the Navier–Stokes Equation

Charles L. Fefferman

Existence and Smoothness of the Navier–Stokes Equation

Charles L. Fefferman

The Euler and Navier–Stokes equations describe the motion of a fluid in $\mathbb{R}^n (n = 2 \text{ or } 3)$. These equations are to be solved for an unknown velocity vector $u(x,t) = (u_i(x,t))_{1\le i\le n} \in \mathbb{R}^n$ and pressure $p(x,t) \in \mathbb{R}$, defined for position $x \in \mathbb{R}^n$ and time $t \ge 0$. We restrict attention here to incompressible fluids filling all of $\mathbb{R}^n$. The *Navier–Stokes* equations are then given by

$$\frac{\partial}{\partial t}u_i + \sum_{j=1}^{n} u_j \frac{\partial u_i}{\partial x_j} = \nu \Delta u_i - \frac{\partial p}{\partial x_i} + f_i(x,t) \qquad (x \in \mathbb{R}^n, t \ge 0), \tag{1}$$

$$\operatorname{div} u = \sum_{i=1}^{n} \frac{\partial u_i}{\partial x_i} = 0 \qquad (x \in \mathbb{R}^n, t \ge 0) \tag{2}$$

with initial conditions

$$u(x,0) = u^\circ(x) \qquad (x \in \mathbb{R}^n). \tag{3}$$

Here, $u^\circ(x)$ is a given, C^∞ divergence-free vector field on $\mathbb{R}^n$, $f_i(x,t)$ are the components of a given, externally applied force (e.g. gravity), ν is a positive coefficient (the viscosity), and $\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian in the space variables. The *Euler equations* are equations (1), (2), (3) with ν set equal to zero.

Equation (1) is just Newton's law $f = ma$ for a fluid element subject to the external force $f = (f_i(x,t))_{1\le i\le n}$ and to the forces arising from pressure and friction. Equation (2) just says that the fluid is incompressible. For physically reasonable solutions, we want to make sure $u(x,t)$ does not grow large as $|x| \to \infty$. Hence, we will restrict attention to forces f and initial conditions u° that satisfy

$$|\partial_x^\alpha u^\circ(x)| \le C_{\alpha K}(1+|x|)^{-K} \qquad \text{on } \mathbb{R}^n, \text{ for any } \alpha \text{ and } K \tag{4}$$

and

$$|\partial_x^\alpha \partial_t^m f(x,t)| \le C_{\alpha m K}(1+|x|+t)^{-K} \quad \text{on } \mathbb{R}^n \times [0,\infty), \text{ for any } \alpha, m, K. \tag{5}$$

We accept a solution of (1), (2), (3) as physically reasonable only if it satisfies

$$p, u \in C^\infty(\mathbb{R}^n \times [0, \infty)) \tag{6}$$

and

$$\int_{\mathbb{R}^n} |u(x,t)|^2 dx < C \quad \text{for all } t \geq 0 \quad \text{(bounded energy)}. \tag{7}$$

Alternatively, to rule out problems at infinity, we may look for spatially periodic solutions of (1), (2), (3). Thus, we assume that $u^\circ(x), f(x,t)$ satisfy

$$u^\circ(x + e_j) = u^\circ(x), \quad f(x + e_j, t) = f(x,t) \quad \text{for } 1 \leq j \leq n \tag{8}$$

($e_j = j^{\text{th}}$ unit vector in $\mathbb{R}^n$).

In place of (4) and (5), we assume that u° is smooth and that

$$|\partial_x^\alpha \partial_t^m f(x,t)| \leq C_{\alpha m K}(1 + |t|)^{-K} \quad \text{on } \mathbb{R}^3 \times [0, \infty), \text{ for any } \alpha, m, K. \tag{9}$$

We then accept a solution of (1), (2), (3) as physically reasonable if it satisfies

$$u(x,t) = u(x + e_j, t) \quad \text{on } \mathbb{R}^3 \times [0, \infty) \text{ for } \quad 1 \leq j \leq n \tag{10}$$

and

$$p, u \in C^\infty(\mathbb{R}^n \times [0, \infty)). \tag{11}$$

A fundamental problem in analysis is to decide whether such smooth, physically reasonable solutions exist for the Navier–Stokes equations. To give reasonable leeway to solvers while retaining the heart of the problem, we ask for a proof of one of the following four statements.

(A) Existence and smoothness of Navier–Stokes solutions on $\mathbb{R}^3$. Take $\nu > 0$ and $n = 3$. Let $u^\circ(x)$ be any smooth, divergence-free vector field satisfying (4). Take $f(x,t)$ to be identically zero. Then there exist smooth functions $p(x,t), u_i(x,t)$ on $\mathbb{R}^3 \times [0, \infty)$ that satisfy (1), (2), (3), (6), (7).

(B) Existence and smoothness of Navier–Stokes solutions in $\mathbb{R}^3/\mathbb{Z}^3$. Take $\nu > 0$ and $n = 3$. Let $u^\circ(x)$ be any smooth, divergence-free vector field satisfying (8); we take $f(x,t)$ to be identically zero. Then there exist smooth functions $p(x,t)$, $u_i(x,t)$ on $\mathbb{R}^3 \times [0, \infty)$ that satisfy (1), (2), (3), (10), (11).

(C) Breakdown of Navier–Stokes solutions on $\mathbb{R}^3$. Take $\nu > 0$ and $n = 3$. Then there exist a smooth, divergence-free vector field $u^\circ(x)$ on $\mathbb{R}^3$ and a smooth $f(x,t)$ on $\mathbb{R}^3 \times [0, \infty)$, satisfying (4), (5), for which there exist no solutions (p, u) of (1), (2), (3), (6), (7) on $\mathbb{R}^3 \times [0, \infty)$.

(D) Breakdown of Navier–Stokes Solutions on $\mathbb{R}^3/\mathbb{Z}^3$. Take $\nu > 0$ and $n = 3$. Then there exist a smooth, divergence-free vector field $u^\circ(x)$ on $\mathbb{R}^3$ and a smooth $f(x,t)$ on $\mathbb{R}^3 \times [0, \infty)$, satisfying (8), (9), for which there exist no solutions (p, u) of (1), (2), (3), (10), (11) on $\mathbb{R}^3 \times [0, \infty)$.

These problems are also open and very important for the Euler equations ($\nu = 0$), although the Euler equation is not on the Clay Institute's list of prize problems.

Let me sketch the main partial results known regarding the Euler and Navier–Stokes equations, and conclude with a few remarks on the importance of the question.

In two dimensions, the analogues of assertions (A) and (B) have been known for a long time (Ladyzhenskaya [**4**]), also for the more difficult case of the Euler equations. This gives no hint about the three-dimensional case, since the main difficulties are absent in two dimensions. In three dimensions, it is known that (A) and (B) hold provided the initial velocity u° satisfies a smallness condition. For initial data $u^\circ(x)$ not assumed to be small, it is known that (A) and (B) hold (also for $\nu = 0$) if the time interval $[0, \infty)$ is replaced by a small time interval $[0, T)$, with T depending on the initial data. For a given initial $u^\circ(x)$, the maximum allowable T is called the "blowup time." Either (A) and (B) hold, or else there is a smooth, divergence-free $u^\circ(x)$ for which (1), (2), (3) have a solution with a finite blowup time. For the Navier–Stokes equations ($\nu > 0$), if there is a solution with a finite blowup time T, then the velocity $(u_i(x,t))_{1\le i\le 3}$ becomes unbounded near the blowup time.

Other unpleasant things are known to happen at the blowup time T, if $T < \infty$. For the Euler equations ($\nu = 0$), if there is a solution (with $f \equiv 0$, say) with finite blowup time T, then the vorticity $\omega(x,t) = \operatorname{curl}_x u(x,t)$ satisfies

$$\int_0^T \left\{ \sup_{x\in\mathbb{R}^3} |\omega(x,t)| \right\} dt = \infty \qquad \text{(Beale–Kato–Majda)},$$

so that the vorticity blows up rapidly.

Many numerical computations appear to exhibit blowup for solutions of the Euler equations, but the extreme numerical instability of the equations makes it very hard to draw reliable conclusions.

The above results are covered very well in the book of Bertozzi and Majda [**1**].

Starting with Leray [**5**], important progress has been made in understanding *weak solutions* of the Navier–Stokes equations. To arrive at the idea of a weak solution of a PDE, one integrates the equation against a test function, and then integrates by parts (formally) to make the derivatives fall on the test function. For instance, if (1) and (2) hold, then, for any smooth

vector field $\theta(x,t) = (\theta_i(x,t))_{1\le i\le n}$ compactly supported in $\mathbb{R}^3 \times (0,\infty)$, a formal integration by parts yields

$$\begin{aligned}(12)\qquad &\iint\limits_{\mathbb{R}^3\times\mathbb{R}} u\cdot\frac{\partial\theta}{\partial t}dxdt - \sum_{ij}\iint\limits_{\mathbb{R}^3\times\mathbb{R}} u_iu_j\frac{\partial\theta_i}{\partial x_j}dxdt\\ &\qquad = \nu\iint\limits_{\mathbb{R}^3\times\mathbb{R}} u\cdot\Delta\theta dxdt + \iint\limits_{\mathbb{R}^3\times\mathbb{R}} f\cdot\theta dxdt - \iint\limits_{\mathbb{R}^3\times\mathbb{R}} p\cdot(\operatorname{div}\theta)dxdt.\end{aligned}$$

Note that (12) makes sense for $u \in L^2$, $f \in L^1$, $p \in L^1$, whereas (1) makes sense only if $u(x,t)$ is twice differentiable in x. Similarly, if $\varphi(x,t)$ is a smooth function, compactly supported in $\mathbb{R}^3 \times (0,\infty)$, then a formal integration by parts and (2) imply

$$(13)\qquad \iint\limits_{\mathbb{R}^3\times\mathbb{R}} u\cdot\bigtriangledown_x\varphi dxdt = 0.$$

A solution of (12), (13) is called a *weak solution* of the Navier–Stokes equations.

A long-established idea in analysis is to prove existence and regularity of solutions of a PDE by first constructing a weak solution, then showing that any weak solution is smooth. This program has been tried for Navier–Stokes with partial success. Leray in [**5**] showed that the Navier–Stokes equations (1), (2), (3) in three space dimensions always have a weak solution (p,u) with suitable growth properties. Uniqueness of weak solutions of the Navier–Stokes equation is *not* known. For the Euler equation, uniqueness of weak solutions is strikingly false. Scheffer [**8**], and, later, Schnirelman [**9**] exhibited weak solutions of the Euler equations on $\mathbb{R}^2 \times \mathbb{R}$ with compact support in spacetime. This corresponds to a fluid that starts from rest at time $t = 0$, begins to move at time $t = 1$ with no outside stimulus, and returns to rest at time $t = 2$, with its motion always confined to a ball $B \subset \mathbb{R}^3$.

Scheffer [**7**] applied ideas from geometric measure theory to prove a partial regularity theorem for suitable weak solutions of the Navier–Stokes equations. Caffarelli–Kohn–Nirenberg [**2**] improved Scheffer's results, and F.-H. Lin [**6**] simplified the proofs of the results in Caffarelli–Kohn–Nirenberg [**2**]. The partial regularity theorem of [**2**], [**6**] concerns a parabolic analogue of the Hausdorff dimension of the singular set of a suitable weak solution of Navier–Stokes. Here, the *singular set* of a weak solution u consists of all points $(x^\circ, t^\circ) \in \mathbb{R}^3 \times \mathbb{R}$ such that u is unbounded in every neighborhood of (x°, t°). (If the force f is smooth, and if (x°, t°) doesn't belong to the

singular set, then it's not hard to show that u can be corrected on a set of measure zero to become smooth in a neighborhood of (x°, t°).)

To define the parabolic analogue of Hausdorff dimension, we use *parabolic cylinders* $Q_r = B_r \times I_r \subset \mathbb{R}^3 \times \mathbb{R}$, where $B_r \subset \mathbb{R}^3$ is a ball of radius r, and $I_r \subset \mathbb{R}$ is an interval of length r^2. Given $E \subset \mathbb{R}^3 \times \mathbb{R}$ and $\delta > 0$, we set

$$\mathcal{P}_{K,\delta}(E) = \inf \left\{ \sum_{i=1}^{\infty} r_i^K : Q_{r_1}, Q_{r_2}, \cdots \text{ cover } E, \text{ and each } r_i < \delta \right\}$$

and then define

$$\mathcal{P}_K(E) = \lim_{\delta \to 0+} \mathcal{P}_{K,\delta}(E).$$

The main results of [**2**], [**6**] may be stated roughly as follows.

THEOREM. *(A) Let u be a weak solution of the Navier–Stokes equations, satisfying suitable growth conditions. Let E be the singular set of u. Then $\mathcal{P}_1(E) = 0$.*

(B) Given a divergence-free vector field $u^\circ(x)$ and a force $f(x,t)$ satisfying (4) and (5), there exists a weak solution of Navier–Stokes (1), (2), (3) satisfying the growth conditions in (A).

In particular, the singular set of u cannot contain a spacetime curve of the form $\{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : x = \phi(t)\}$. This is the best partial regularity theorem known so far for the Navier–Stokes equation. It appears to be very hard to go further.

Let me end with a few words about the significance of the problems posed here. Fluids are important and hard to understand. There are many fascinating problems and conjectures about the behavior of solutions of the Euler and Navier–Stokes equations. (See, for instance, Bertozzi–Majda [**1**] or Constantin [**3**].) Since we don't even know whether these solutions exist, our understanding is at a very primitive level. Standard methods from PDE appear inadequate to settle the problem. Instead, we probably need some deep, new ideas.

Bibliography

[1] A. Bertozzi and A. Majda, *Vorticity and Incompressible Flows*, Cambridge U. Press, Cambridge, 2002.

[2] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier–Stokes equations*, Comm. Pure & Appl. Math. **35** (1982), 771–831.

[3] P. Constantin, *Some open problems and research directions in the mathematical study of fluid dynamics*, in Mathematics Unlimited–2001 and Beyond, Springer Verlag, Berlin, 2001, 353–360.

[4] O. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flows* (2nd edition), Gordon and Breach, New York, 1969.

[5] J. Leray, *Sur le mouvement d'un liquide visquex emplissent l'espace*, Acta Math. J. **63** (1934), 193–248

[6] F.-H. Lin, *A new proof of the Caffarelli–Kohn–Nirenberg theorem*, Comm. Pure. & Appl. Math. **51** (1998), 241–257
[7] V. Scheffer, *Turbulence and Hausdorff dimension*, in Turbulence and the Navier–Stokes Equations, Lecture Notes in Math. **565**, Springer Verlag, Berlin, 1976, 94–112.
[8] V. Scheffer, *An inviscid flow with compact support in spacetime*, J. Geom. Analysis **3** (1993), 343–401
[9] A. Shnirelman, *On the nonuniqueness of weak solutions of the Euler equation*, Comm. Pure & Appl. Math. **50** (1997), 1260–1286

The Poincaré Conjecture

John Milnor

The Poincaré Conjecture

John Milnor

1. Introduction

The topology of two-dimensional manifolds or *surfaces* was well understood in the 19th century. In fact there is a simple list of all possible smooth compact orientable surfaces. Any such surface has a well-defined *genus* $g \geq 0$, which can be described intuitively as the number of holes; and two such surfaces can be put into a smooth one-to-one correspondence with each other if and only if they have the same genus.[1] The corresponding question

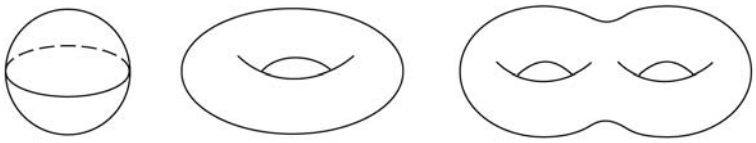


Figure 1. Sketches of smooth surfaces of genus 0, 1, and 2.

in higher dimensions is much more difficult. Henri Poincaré was perhaps the first to try to make a similar study of three-dimensional manifolds. The most basic example of such a manifold is the three-dimensional *unit sphere*, that is, the locus of all points (x, y, z, w) in four-dimensional Euclidean space which have distance exactly 1 from the origin: $x^2 + y^2 + z^2 + w^2 = 1$. He noted that a distinguishing feature of the two-dimensional sphere is that every simple closed curve in the sphere can be deformed continuously to a point without leaving the sphere. In 1904, he asked a corresponding question in dimension 3. In more modern language, it can be phrased as follows:[2]

[1] For definitions and other background material, see, for example, [**21**] or [**29**], as well as [**48**].

[2] See [**36**, pages 498 and 370]. To Poincaré, manifolds were always smooth or polyhedral, so that his term "homeomorphism" referred to a smooth or piecewise linear homeomorphism.

QUESTION. *If a compact three-dimensional manifold M^3 has the property that every simple closed curve within the manifold can be deformed continuously to a point, does it follow that M^3 is homeomorphic to the sphere S^3?*

He commented, with considerable foresight, "*Mais cette question nous entraînerait trop loin*". Since then, the hypothesis that every simply connected closed 3-manifold is homeomorphic to the 3-sphere has been known as the Poincaré Conjecture. It has inspired topologists ever since, and attempts to prove it have led to many advances in our understanding of the topology of manifolds.

2. Early Missteps

From the first, the apparently simple nature of this statement has led mathematicians to overreach. Four years earlier, in 1900, Poincaré himself had been the first to err, stating a false theorem that can be phrased as follows.

FALSE THEOREM. *Every compact polyhedral manifold with the homology of an n-dimensional sphere is actually homeomorphic to the n-dimensional sphere.*

But his 1904 paper provided a beautiful counterexample to this claim, based on the concept of *fundamental group*, which he had introduced earlier (see [**36**, pp. 189–192 and 193–288]). This example can be described geometrically as follows. Consider all possible regular icosahedra inscribed in the two-dimensional unit sphere. In order to specify one particular icosahedron in this family, we must provide three parameters. For example, two parameters are needed to specify a single vertex on the sphere, and then another parameter to specify the direction to a neighboring vertex. *Thus each such icosahedron can be considered as a single "point" in the three-dimensional manifold M^3 consisting of all such icosahedra.*[3] This manifold meets Poincaré's preliminary criterion: By the methods of homology theory, it cannot be distinguished from the three-dimensional sphere. However, he could prove that it is not a sphere by constructing a simple closed curve that cannot be deformed to a point within M^3. The construction is not difficult: Choose some representative icosahedron and consider its images under rotation about one vertex through angles $0 \le \theta \le 2\pi/5$. This defines a simple closed curve in M^3 that cannot be deformed to a point.

[3]In more technical language, this M^3 can be defined as the coset space $\mathrm{SO}(3)/\mathrm{I}_{60}$ where $\mathrm{SO}(3)$ is the group of all rotations of Euclidean 3-space and where I_{60} is the subgroup consisting of the 60 rotations that carry a standard icosahedron to itself. The fundamental group $\pi_1(M^3)$, consisting of all homotopy classes of loops from a point to itself within M^3, is a perfect group of order 120.

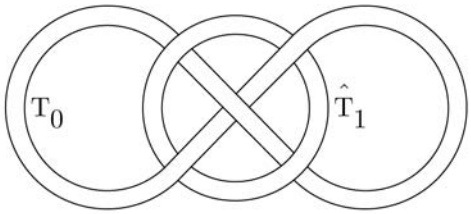


FIGURE 2. The Whitehead link

The next important false theorem was by Henry Whitehead in 1934 [**52**]. As part of a purported proof of the Poincaré Conjecture, he claimed the sharper statement that every open three-dimensional manifold that is *contractible* (that can be continuously deformed to a point) is homeomorphic to Euclidean space. Following in Poincaré's footsteps, he then substantially increased our understanding of the topology of manifolds by discovering a counterexample to his own theorem. His counterexample can be briefly described as follows. Start with two disjoint solid tori T_0 and $\widehat{T}_1$ in the 3-sphere that are embedded as shown in Figure 2, so that each one individually is unknotted, but so that the two are linked together with linking number zero. Since $\widehat{T}_1$ is unknotted, its complement $T_1 = S^3 \smallsetminus \text{interior}(\widehat{T}_1)$ is another unknotted solid torus that contains T_0. Choose a homeomorphism h of the 3-sphere that maps T_0 onto this larger solid torus T_1. Then we can inductively construct solid tori

$$T_0 \subset T_1 \subset T_2 \subset \cdots$$

in S^3 by setting $T_{j+1} = h(T_j)$. The union $M^3 = \bigcup T_j$ of this increasing sequence is the required Whitehead counterexample, a contractible manifold that is not homeomorphic to Euclidean space. To see that $\pi_1(M^3) = 0$, note that every closed loop in T_0 can be shrunk to a point (after perhaps crossing through itself) within the larger solid torus T_1. But every closed loop in M^3 must be contained in some T_j, and hence can be shrunk to a point within $T_{j+1} \subset M^3$. On the other hand, M^3 is not homeomorphic to Euclidean 3-space since, if $K \subset M^3$ is any compact subset large enough to contain T_0, one can prove that the difference set $M^3 \smallsetminus K$ is not simply connected.

Since this time, many false proofs of the Poincaré Conjecture have been proposed, some of them relying on errors that are rather subtle and difficult to detect. For a delightful presentation of some of the pitfalls of three-dimensional topology, see [**4**].

3. Higher Dimensions

The late 1950s and early 1960s saw an avalanche of progress with the discovery that higher-dimensional manifolds are actually easier to work with

than three-dimensional ones. One reason for this is the following: The fundamental group plays an important role in all dimensions even when it is trivial, and relations between generators of the fundamental group correspond to two-dimensional disks, mapped into the manifold. In dimension 5 or greater, such disks can be put into general position so that they are disjoint from each other, with no self-intersections, but in dimension 3 or 4 it may not be possible to avoid intersections, leading to serious difficulties.

Stephen Smale announced a proof of the Poincaré Conjecture in high dimensions in 1960 [**41**]. He was quickly followed by John Stallings, who used a completely different method [**43**], and by Andrew Wallace, who had been working along lines quite similar to those of Smale [**51**].

Let me first describe the Stallings result, which has a weaker hypothesis and easier proof, but also a weaker conclusion. He assumed that the dimension is seven or more, but Christopher Zeeman later extended his argument to dimensions 5 and 6 [**54**].

Stallings–Zeeman Theorem. *If M^n is a finite simplicial complex of dimension $n \geq 5$ that has the homotopy type*[4] *of the sphere S^n and is locally piecewise linearly homeomorphic to the Euclidean space $\mathbb{R}^n$, then M^n is homeomorphic to S^n under a homeomorphism that is piecewise linear except at a single point. In other words, the complement $M^n \setminus (point)$ is piecewise linearly homeomorphic to $\mathbb{R}^n$.*

The method of proof consists of pushing all of the difficulties off toward a single point; hence there can be no control near that point.

The Smale proof, and the closely related proof given shortly afterward by Wallace, depended rather on differentiable methods, building a manifold up inductively, starting with an n-dimensional ball, by successively adding handles. Here a *k-handle* can be added to a manifold M^n with boundary by first attaching a k-dimensional cell, using an attaching homeomorphism from the $(k-1)$-dimensional boundary sphere into the boundary of M^n, and then thickening and smoothing corners so as to obtain a larger manifold with boundary. The proof is carried out by rearranging and canceling such handles. (Compare the presentation in [**24**].)

Smale Theorem. *If M^n is a differentiable homotopy sphere of dimension $n \geq 5$, then M^n is homeomorphic to S^n. In fact, M^n is diffeomorphic to a manifold obtained by gluing together the boundaries of two closed n-balls under a suitable diffeomorphism.*

[4]In order to check that a manifold M^n has the same homotopy type as the sphere S^n, we must check not only that it is simply connected, $\pi_1(M^n) = 0$, but also that it has the same homology as the sphere. The example of the product $S^2 \times S^2$ shows that it is not enough to assume that $\pi_1(M^n) = 0$ when $n > 3$.

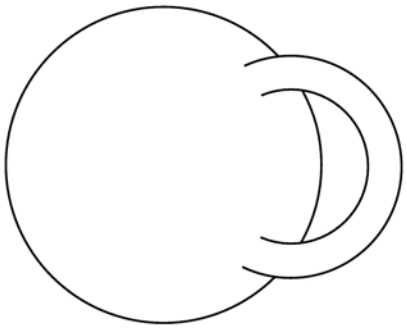


FIGURE 3. A three-dimensional ball with a 1-handle attached

This was also proved by Wallace, at least for $n \geq 6$. (It should be noted that the five-dimensional case is particularly difficult.)

The much more difficult four-dimensional case had to wait twenty years, for the work of Michael Freedman [**8**]. Here the differentiable methods used by Smale and Wallace and the piecewise linear methods used by Stallings and Zeeman do not work at all. Freedman used wildly non-differentiable methods, not only to prove the four-dimensional Poincaré Conjecture for topological manifolds, but also to give a complete classification of all closed simply connected topological 4-manifolds. The integral cohomology group H^2 of such a manifold is free abelian. Freedman needed just two invariants: The cup product $\beta : H^2 \otimes H^2 \to H^4 \cong \mathbb{Z}$ is a symmetric bilinear form with determinant ± 1, while the *Kirby–Siebenmann invariant* κ is an integer mod 2 that vanishes if and only if the product manifold $M^4 \times \mathbb{R}$ can be given a differentiable structure.

FREEDMAN THEOREM. *Two closed simply connected 4-manifolds are homeomorphic if and only if they have the same bilinear form β and the same Kirby–Siebenmann invariant κ. Any β can be realized by such a manifold. If $\beta(x \otimes x)$ is odd for some $x \in H^2$, then either value of κ can be realized also. However, if $\beta(x \otimes x)$ is always even, then κ is determined by β, being congruent to one eighth of the signature of β.*

In particular, if M^4 is a homotopy sphere, then $H^2 = 0$ and $\kappa = 0$, so M^4 is homeomorphic to S^4. It should be noted that the piecewise linear or differentiable theories in dimension 4 are much more difficult. It is not known whether every smooth homotopy 4-sphere is diffeomorphic to S^4; it is not known which 4-manifolds with $\kappa = 0$ actually possess differentiable structures; and it is not known when this structure is essentially unique. The major results on these questions are due to Simon Donaldson [**7**]. As

one indication of the complications, Freedman showed, using Donaldson's work, that $\mathbb{R}^4$ admits uncountably many inequivalent differentiable structures. (Compare [**12**].)

In dimension 3, the discrepancies between topological, piecewise linear, and differentiable theories disappear (see [**18**], [**28**], and [**26**]). However, difficulties with the fundamental group become severe.

4. The Thurston Geometrization Conjecture

In the two-dimensional case, each smooth compact surface can be given a beautiful geometrical structure, as a round sphere in the genus zero case, as a flat torus in the genus 1 case, and as a surface of constant negative curvature when the genus is 2 or more. A far-reaching conjecture by William Thurston in 1983 claims that something similar is true in dimension 3 [**46**]. This conjecture asserts that every compact orientable three-dimensional manifold can be cut up along 2-spheres and tori so as to decompose into essentially unique pieces, each of which has a simple geometrical structure. There are eight possible three-dimensional geometries in Thurston's program. Six of these are now well understood,[5] and there has been a great deal of progress with the geometry of constant negative curvature.[6] The eighth geometry, however, corresponding to constant positive curvature, remains largely untouched. For this geometry, we have the following extension of the Poincaré Conjecture.

Thurston Elliptization Conjecture. *Every closed 3-manifold with finite fundamental group has a metric of constant positive curvature and hence is homeomorphic to a quotient S^3/Γ, where $\Gamma \subset \mathrm{SO}(4)$ is a finite group of rotations that acts freely on S^3.*

The Poincaré Conjecture corresponds to the special case where the group $\Gamma \cong \pi_1(M^3)$ is trivial. The possible subgroups $\Gamma \subset \mathrm{SO}(4)$ were classified long ago by [**19**] (compare [**23**]), but this conjecture remains wide open.

5. Approaches through Differential Geometry and Differential Equations[7]

In recent years there have been several attacks on the geometrization problem (and hence on the Poincaré Conjecture) based on a study of the geometry of the infinite dimensional space consisting of all Riemannian metrics on a given smooth three-dimensional manifold.

[5]See, for example, [**13**], [**3**], [**38, 39, 40**], [**49**], [**9**], and [**6**].

[6]See [**44**], [**27**], [**47**], [**22**], and [**30**]. The pioneering papers by [**14**] and [**50**] provided the basis for much of this work.

[7]Added in 2004

By definition, the length of a path γ on a Riemannian manifold is computed, in terms of the *metric tensor* g_{ij}, as the integral

$$\int_\gamma ds = \int_\gamma \sqrt{\sum g_{ij} dx^i dx^j}\,.$$

From the first and second derivatives of this metric tensor, one can compute the *Ricci curvature tensor* R_{ij}, and the *scalar curvature* R. (As an example, for the flat Euclidean space one gets $R_{ij} = R = 0$, while for a round three-dimensional sphere of radius r, one gets Ricci curvature $R_{ij} = 2g_{ij}/r^2$ and scalar curvature $R = 6/r^2$.)

One approach by Michael Anderson, based on ideas of Hidehiko Yamabe [**53**], studies the *total scalar curvature* $\iiint_{M^3} R\,dV$ as a functional on the space of all smooth unit volume Riemannian metrics. The critical points of this functional are the metrics of constant curvature (see [**1**]).

A different approach, initiated by Richard Hamilton studies the *Ricci flow* [**15, 16, 17**], that is, the solutions to the differential equation

$$\frac{dg_{ij}}{dt} = -2R_{ij}\,.$$

In other words, the metric is required to change with time so that distances decrease in directions of positive curvature. This is essentially a parabolic differential equationa and behaves much like the heat equation studied by physicists: If we heat one end of a cold rod, then the heat will gradually flow throughout the rod until it attains an even temperature. Similarly, a naive hope for 3-manifolds with finite fundamental group might have been that, under the Ricci flow, positive curvature would tend to spread out until, in the limit (after rescaling to constant size), the manifold would attain constant curvature. If we start with a 3-manifold of positive Ricci curvature, Hamilton was able to carry out this program and construct a metric of constant curvature, thus solving a very special case of the Elliptization Conjecture. However, in the general case, there are very serious difficulties, since this flow may tend toward singularities.[8]

I want to thank many mathematicians who helped me with this report.

May 2000, revised June 2004

[8]Grisha Perelman, in St. Petersburg, has posted three preprints on arXiv.org which go a long way toward resolving these difficulties, and in fact claim to prove the full geometrization conjecture [**32, 33, 34**]. These preprints have generated a great deal of interest. (Compare [**2**] and [**25**], as well as the website http://www.math.lsa.umich.edu/research/ricciflow/perelman.html organized by B. Kleiner and J. Lott.) However, full details have not appeared.

Bibliography

[1] M.T. Anderson, *Scalar curvature, metric degenerations and the static vacuum Einstein equations on 3-manifolds*, Geom. Funct. Anal. **9** (1999), 855–963 and **11** (2001) 273–381. See also: *Scalar curvature and the existence of geometric structures on 3-manifolds*, J. reine angew. Math. **553** (2002), 125–182 and **563** (2003), 115–195.

[2] M.T. Anderson, *Geometrization of 3-manifolds via the Ricci flow*, Notices AMS **51** (2004), 184–193.

[3] L. Auslander and F.E.A. Johnson, *On a conjecture of C.T.C. Wall*, J. Lond. Math. Soc. **14** (1976), 331–332.

[4] R.H. Bing, *Some aspects of the topology of 3-manifolds related to the Poincaré conjecture*, in Lectures on Modern Mathematics II (T. L. Saaty, ed.), Wiley, New York, 1964.

[5] J. Birman, *Poincaré's conjecture and the homeotopy group of a closed, orientable* 2*-manifold*, J. Austral. Math. Soc. **17** (1974), 214–221.

[6] A. Casson and D. Jungreis, *Convergence groups and Seifert fibered* 3*-manifolds*, Invent. Math. 118 (1994), 441–456.

[7] S.K. Donaldson, *Self-dual connections and the topology of smooth* 4*-manifolds*, Bull. Amer. Math. Soc. **8** (1983), 81–83.

[8] M.H. Freedman, *The topology of four-dimensional manifolds*, J. Diff. Geom. **17** (1982), 357–453.

[9] D. Gabai, *Convergence groups are Fuchsian groups*, Ann. Math. 136 (1992), 447–510.

[10] D. Gabai, *Valentin Poenaru's program for the Poincaré conjecture*, in Geometry, topology, & physics, Conf. Proc. Lecture Notes Geom. Topology, VI, Internat. Press, Cambridge, MA, 1995, 139–166.

[11] D. Gillman and D. Rolfsen, *The Zeeman conjecture for standard spines is equivalent to the Poincaré conjecture*, Topology **22** (1983), 315–323.

[12] R. Gompf, *An exotic menagerie*, J. Differential Geom. **37** (1993) 199–223.

[13] C. Gordon and W. Heil, *Cyclic normal subgroups of fundamental groups of 3-manifolds*, Topology **14** (1975), 305–309.

[14] W. Haken, *Über das Homöomorphieproblem der 3-Mannigfaltigkeiten I*, Math. Z. **80** (1962), 89–120.

[15] R.S. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom. **17** (1982), 255–306.

[16] R.S. Hamilton, *The formation of singularities in the Ricci flow*, in Surveys in differential geometry, Vol. II (Cambridge, MA, 1993), Internat. Press, Cambridge, MA, 1995, 7–136.

[17] R.S. Hamilton, *Non-singular solutions of the Ricci flow on three-manifolds* Comm. Anal. Geom. **7** (1999), 695–729.

[18] M. Hirsch, *Obstruction theories for smoothing manifolds and maps*, Bull. Amer. Math. Soc. **69** (1963), 352-356.

[19] H. Hopf, *Zum Clifford–Kleinschen Raumproblem*, Math. Ann. **95** (1925-26) 313-319.

[20] W. Jakobsche, *The Bing-Borsuk conjecture is stronger than the Poincaré conjecture*, Fund. Math. **106** (1980), 127–134.

[21] W.S. Massey, *Algebraic Topology: An Introduction*, Harcourt Brace, New York, 1967; Springer, New York 1977; or *A Basic Course in Algebraic Topology*, Springer, New York, 1991.

[22] C. McMullen, *Riemann surfaces and geometrization of 3-manifolds*, Bull. Amer. Math. Soc. **27** (1992), 207–216.

[23] J. Milnor, *Groups which act on S^n without fixed points*, Amer. J. Math. **79** (1957), 623–630.

[24] J. Milnor (with L. Siebenmann and J. Sondow), *Lectures on the h-Cobordism Theorem*, Princeton Math. Notes, Princeton University Press, Princeton, 1965.

[25] J. Milnor, *Towards the Poincaré conjecture and the classification of 3-manifolds*, Notices AMS **50** (2003), 1226–1233.
[26] E.E. Moise, *Geometric Topology in Dimensions* 2 *and* 3, Springer, New York, 1977.
[27] J. Morgan, *On Thurston's uniformization theorem for three-dimensional manifolds*, in The Smith Conjecture (H. Bass and J. Morgan, eds.), Pure and Appl. Math. 112, Academic Press, New York, 1984, 37–125.
[28] J. Munkres, *Obstructions to the smoothing of piecewise-differentiable homeomorphisms,* Ann. Math. **72** (1960), 521–554.
[29] J. Munkres, *Topology: A First Course*, Prentice–Hall, Englewood Cliffs, NJ, 1975.
[30] J.-P. Otal, *The hyperbolization theorem for fibered 3-manifolds*, translated from the 1996 French original by Leslie D. Kay, SMF/AMS Texts and Monographs 7, American Mathematical Society, Providence, RI; Société Mathatique de France, Paris, 2001.
[31] C. Papakyriakopoulos, *A reduction of the Poincaré conjecture to group theoretic conjectures*, Ann. Math. **77** (1963), 250–305.
[32] G. Perelman, *The entropy formula for the Ricci flow and its geometric applications*, arXiv: math.DG/0211159v1, 11 Nov 2002.
[33] G. Perelman, *Ricci flow with surgery on three-manifolds*, arXiv: math.DG/0303109, 10 Mar 2003.
[34] G. Perelman, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, arXiv: math.DG/0307245, 17 Jul 2003.
[35] V. Poénaru, *A program for the Poincaré conjecture and some of its ramifications*, in Topics in low-dimensional topology (University Park, PA, 1996), World Sci. Publishing, River Edge, NJ, 1999, 65–88.
[36] H. Poincaré, Œuvres, Tome VI, Gauthier–Villars, Paris, 1953.
[37] C. Rourke, *Algorithms to disprove the Poincaré conjecture*, Turkish J. Math. **21** (1997), 99–110.
[38] P. Scott, *A new proof of the annulus and torus theorems*, Amer. J. Math. **102** (1980), 241–277.
[39] P. Scott, *There are no fake Seifert fibre spaces with infinite* π_1, Ann. Math. **117** (1983), 35–70.
[40] P. Scott, *The geometries of 3-manifolds*, Bull. Lond. Math. Soc. **15** (1983), 401–487.
[41] S. Smale, *Generalized Poincaré's conjecture in dimensions greater than four*, Ann. Math. **74** (1961), 391–406. (See also: Bull. Amer. Math. Soc. **66** (1960), 373–375.)
[42] S. Smale, *The story of the higher dimensional Poincaré conjecture* (*What actually happened on the beaches of Rio*), Math. Intelligencer **12**, no. 2 (1990), 44–51.
[43] J. Stallings, *Polyhedral homotopy spheres,* Bull. Amer. Math. Soc. **66** (1960), 485–488.
[44] D. Sullivan, *Travaux de Thurston sur les groupes quasi-fuchsiens et sur les variétés hyperboliques de dimension 3 fibrées sur le cercle*, Sém. Bourbaki 554, Lecture Notes Math. **842**, Springer, New York, 1981.
[45] T.L. Thickstun, *Open acyclic* 3*-manifolds, a loop theorem and the Poincaré conjecture*, Bull. Amer. Math. Soc. (N.S.) **4** (1981), 192–194.
[46] W.P. Thurston, *Three dimensional manifolds, Kleinian groups and hyperbolic geometry*, in The Mathematical heritage of Henri Poincaré, Proc. Symp. Pure Math. **39** (1983), Part 1. (Also in Bull. Amer. Math. Soc. **6** (1982), 357–381.)
[47] W.P. Thurston, *Hyperbolic structures on 3-manifolds, I, deformation of acyclic manifolds*, Ann. Math. **124** (1986), 203–246
[48] W.P. Thurston, *Three-Dimensional Geometry and Topology*, Vol. 1, ed. by Silvio Levy, Princeton Mathematical Series **35**, Princeton University Press, Princeton, 1997.
[49] P. Tukia, *Homeomorphic conjugates of Fuchsian groups*, J. Reine Angew. Math. 391 (1988), 1–54.
[50] F. Waldhausen, *On irreducible 3-manifolds which are sufficiently large*, Ann. Math. 87 (1968), 56–88.

[51] A. Wallace, *Modifications and cobounding manifolds, II*, J. Math. Mech **10** (1961), 773–809.
[52] J.H.C. Whitehead, *Mathematical Works*, Volume II, Pergamon Press, New York, 1962. (See pages 21-50.)
[53] H. Yamabe, *On a deformation of Riemannian structures on compact manifolds*, Osaka Math. J. **12** (1960), 21–37.
[54] E.C. Zeeman, *The Poincaré conjecture for* $n \geq 5$, in Topology of 3-Manifolds and Related Topics Prentice–Hall, Englewood Cliffs, NJ, 1962, 198–204. (See also Bull. Amer. Math. Soc. **67** (1961), 270.)

(Note: For a representative collection of attacks on the Poincaré Conjecture, see [**31**], [**5**], [**20**], [**45**], [**11**], [**10**], [**37**], and [**35**].)

The P versus NP Problem

Stephen Cook

The P versus NP Problem

Stephen Cook

1. Statement of the Problem

The **P** versus **NP** problem is to determine whether every language accepted by some nondeterministic algorithm in polynomial time is also accepted by some (deterministic) algorithm in polynomial time. To define the problem precisely it is necessary to give a formal model of a computer. The standard computer model in computability theory is the Turing machine, introduced by Alan Turing in 1936 [**37**]. Although the model was introduced before physical computers were built, it nevertheless continues to be accepted as the proper computer model for the purpose of defining the notion of *computable function*.

Informally the class **P** is the class of decision problems solvable by some algorithm within a number of steps bounded by some fixed polynomial in the length of the input. Turing was not concerned with the efficiency of his machines, rather his concern was whether they can simulate arbitrary algorithms given sufficient time. It turns out, however, Turing machines can generally simulate more efficient computer models (for example, machines equipped with many tapes or an unbounded random access memory) by at most squaring or cubing the computation time. Thus **P** is a robust class and has equivalent definitions over a large class of computer models. Here we follow standard practice and define the class **P** in terms of Turing machines.

Formally the elements of the class **P** are languages. Let Σ be a finite alphabet (that is, a finite nonempty set) with at least two elements, and let Σ^* be the set of finite strings over Σ. Then a *language over* Σ is a subset L of Σ^*. Each Turing machine M has an associated *input alphabet* Σ. For each string w in Σ^* there is a computation associated with M with input w. (The notions of Turing machine and computation are defined formally in the appendix.) We say that M *accepts* w if this computation terminates in the accepting state. Note that M fails to accept w either if this computation ends in the rejecting state, or if the computation fails to terminate. The *language accepted by* M, denoted $L(M)$, has associated alphabet Σ and is

defined by

$$L(M) = \{w \in \Sigma^* \mid M \text{ accepts } w\}.$$

We denote by $t_M(w)$ the number of steps in the computation of M on input w (see the appendix). If this computation never halts, then $t_M(w) = \infty$. For $n \in \mathbb{N}$ we denote by $T_M(n)$ the *worst case run time* of M; that is,

$$T_M(n) = \max\{t_M(w) \mid w \in \Sigma^n\},$$

where Σ^n is the set of all strings over Σ of length n. We say that M *runs in polynomial time* if there exists k such that for all n, $T_M(n) \leq n^k + k$. Now we define the class **P** of languages by

$$\mathbf{P} = \{L \mid L = L(M) \text{ for some Turing machine } M \text{ that runs in polynomial time}\}.$$

The notation **NP** stands for "nondeterministic polynomial time", since originally **NP** was defined in terms of nondeterministic machines (that is, machines that have more than one possible move from a given configuration). However, now it is customary to give an equivalent definition using the notion of a *checking relation*, which is simply a binary relation $R \subseteq \Sigma^* \times \Sigma_1^*$ for some finite alphabets Σ and Σ_1. We associate with each such relation R a language L_R over $\Sigma \cup \Sigma_1 \cup \{\#\}$ defined by

$$L_R = \{w\#y \mid R(w, y)\}$$

where the symbol $\#$ is not in Σ. We say that R is *polynomial-time* iff $L_R \in \mathbf{P}$.

Now we define the class **NP** of languages by the condition that a language L over Σ is in **NP** iff there is $k \in \mathbb{N}$ and a polynomial-time checking relation R such that for all $w \in \Sigma^*$,

$$w \in L \iff \exists y(|y| \leq |w|^k \text{ and } R(w, y)),$$

where $|w|$ and $|y|$ denote the lengths of w and y, respectively.

Problem Statement. *Does* $\mathbf{P} = \mathbf{NP}$*?*

It is easy to see that the answer is independent of the size of the alphabet Σ (we assume $|\Sigma| \geq 2$), since strings over an alphabet of any fixed size can be efficiently coded by strings over a binary alphabet. (For $|\Sigma| = 1$ the problem is still open, although it is possible that $\mathbf{P} = \mathbf{NP}$ in this case but not in the general case.)

It is trivial to show that $\mathbf{P} \subseteq \mathbf{NP}$, since for each language L over Σ, if $L \in \mathbf{P}$ then we can define the polynomial-time checking relation $R \subseteq \Sigma^* \cup \Sigma^*$ by

$$R(w, y) \iff w \in L$$

for all $w, y \in \Sigma^*$.

Here are two simple examples, using decimal notation to code natural numbers: The set of perfect squares is in **P**, since Newton's method can be used to efficiently approximate square roots. The set of composite numbers is in **NP**, where (denoting the decimal notation for a natural number c by $\overline{c}$) the associated polynomial time checking relation R is given by

$$R(\overline{a}, \overline{b}) \iff 1 < b < a \text{ and } b|a. \tag{1}$$

(Recently it was shown that in fact the set of composite numbers is also in **P** [**1**], answering a longstanding open question.)

2. History and Importance

The importance of the **P** vs **NP** question stems from the successful theories of **NP**-completeness and complexity-based cryptography, as well as the potentially stunning practical consequences of a constructive proof of **P** = **NP**.

The theory of **NP**-completeness has its roots in computability theory, which originated in the work of Turing, Church, Gödel, and others in the 1930s. The computability precursors of the classes **P** and **NP** are the classes of decidable and c.e. (computably enumerable) languages, respectively. We say that a language L is c.e. (or *semi-decidable*) iff $L = L(M)$ for some Turing machine M. We say that L is *decidable* iff $L = L(M)$ for some Turing machine M that satisfies the condition that M halts on all input strings w. There is an equivalent definition of c.e. that brings out its analogy with **NP**, namely L is c.e. iff there is a computable "checking relation" $R(x, y)$ such that $L = \{x \mid \exists y R(x, y)\}$.

Using the notation $\langle M \rangle$ to denote a string describing a Turing machine M, we define the Halting Problem HP as follows:

$$HP = \{\langle M \rangle \mid M \text{ is a Turing machine that halts on input } \langle M \rangle\}.$$

Turing used a simple diagonal argument to show that HP is not decidable. On the other hand, it is not hard to show that HP is c.e.

Of central importance in computability theory is the notion of reducibility, which Turing defined roughly as follows: A language L_1 is *Turing reducible* to a language L_2 iff there is an oracle Turing machine M that accepts L_1, where M is allowed to make membership queries of the form $x \in L_2$, which are correctly answered by an "oracle" for L_2. Later, the more restricted notion of many-one reducibility ($\leq_m$) was introduced and defined as follows.

Definition 1. Suppose that L_i is a language over Σ_i, $i = 1, 2$. Then $L_1 \leq_m L_2$ iff there is a (total) computable function $f : \Sigma_1^* \to \Sigma_2^*$ such that $x \in L_1 \iff f(x) \in L_2$, for all $x \in \Sigma_1^*$.

It is easy to see that if $L_1 \leq_m L_2$ and L_2 is decidable, then L_1 is decidable. This fact provides an important tool for showing undecidability; for example, if $HP \leq_m L$, then L is undecidable.

The notion of **NP**-complete is based on the following notion from computability theory:

DEFINITION 2. A language L is c.e.-*complete* iff L is c.e., and $L' \leq_m L$ for every c.e. language L'.

It is easy to show that HP is c.e.-complete. It turns out that most "natural" undecidable c.e. languages are, in fact, c.e.-complete. Since $\leq_m$ is transitive, to show that a c.e. language L is c.e.-complete it suffices to show that $HP \leq_m L$.

The notion of polynomial-time computation was introduced in the 1960s by Cobham [**8**] and Edmonds [**13**] as part of the early development of computational complexity theory (although earlier von Neumann [**38**], in 1953, distinguished between polynomial-time and exponential-time algorithms). Edmonds called polynomial-time algorithms "good algorithms" and linked them to tractable algorithms.

It has now become standard in complexity theory to identify polynomial-time with feasible, and here we digress to discuss this point. It is of course not literally true that every polynomial-time algorithm can be feasibly executed on small inputs; for example, a computer program requiring n^{100} steps could never be executed on an input even as small as $n = 10$. Here is a more defensible statement (see [**10**]):

FEASIBILITY THESIS. *A natural problem has a feasible algorithm iff it has a polynomial-time algorithm.*

Examples of natural problems that have both feasible and polynomial-time algorithms abound: Integer arithmetic, linear algebra, network flow, linear programming, many graph problems (connectivity, shortest path, minimum spanning tree, bipartite matching), etc. On the other hand, the deep results of Robertson and Seymour [**29**] provide a potential source of counterexamples to the thesis: They prove that every minor-closed family of graphs can be recognized in polynomial time (in fact, in time $O(n^3)$), but the algorithms supplied by their method have such huge constants that they are not feasible. However, each potential counterexample can be removed by finding a feasible algorithm for it. For example, a feasible recognition algorithm is known for the class of planar graphs, but none is currently known for the class of graphs embeddable in $\mathbb{R}^3$ with no two cycles linked. (Both examples are minor-closed families.) Of course the words "natural" and "feasible"in the thesis above should be explained; generally we do not

consider a class with a parameter as natural, such as the set of graphs embeddable on a surface of genus k, $k > 1$.

We mention two concerns related to the "only if" direction of the thesis. The first comes from randomized algorithms. We discuss at the end of Section 3 the possibility that a source of random bits might be used to greatly reduce the recognition time required for some language. Note, however, that it is not clear whether a truly random source exists in nature. The second concern comes from quantum computers. This computer model incorporates the idea of superposition of states from quantum mechanics and allows a potential exponential speed-up of some computations over Turing machines. For example, Shor [**32**] has shown that some quantum computer algorithm is able to factor integers in polynomial time, but no polynomial-time integer-factoring algorithm is known for Turing machines. Physicists have so far been unable to build a quantum computer that can handle more than a half-dozen bits, so this threat to the feasibility thesis is hypothetical at present.

Returning to the historical treatment of complexity theory, in 1971 the present author [**9**] introduced a notion of **NP**-completeness as a polynomial-time analog of c.e.-completeness, except that the reduction used was a polynomial-time analog of Turing reducibility rather than of many-one reducibility. The main results in [**9**] are that several natural problems, including Satisfiability and 3-SAT (defined below) and subgraph isomorphism are **NP**-complete. A year later Karp [**21**] used these completeness results to show that 20 other natural problems are **NP**-complete, thus forcefully demonstrating the importance of the subject. Karp also introduced the now standard notation **P** and **NP** and redefined **NP**-completeness using the polynomial-time analog of many-one reducibility, a definition that has become standard. Meanwhile Levin [**23**], independently of Cook and Karp, defined the notion of "universal search problem", similar to the **NP**-complete problem, and gave six examples, including Satisfiability.

The standard definitions concerning **NP**-completeness are close analogs of Definitions 1 and 2 above.

DEFINITION 3. Suppose that L_i is a language over Σ_i, $i = 1, 2$. Then $L_1 \leq_p L_2$ (L_1 is p-reducible to L_2) iff there is a polynomial-time computable function $f : \Sigma_1^* \to \Sigma_2^*$ such that $x \in L_1 \iff f(x) \in L_2$, for all $x \in \Sigma_1^*$.

DEFINITION 4. A language L is **NP**-*complete* iff L is in **NP**, and $L' \leq_p L$ for every language L' in **NP**.

The following proposition is easy to prove: Part (b) uses the transitivity of $\leq_p$, and part (c) follows from part (a).

PROPOSITION 1. *(a) If $L_1 \leq_p L_2$ and $L_2 \in \mathbf{P}$, then $L_1 \in \mathbf{P}$.*

(b) If L_1 is **NP***-complete, $L_2 \in$* **NP***, and $L_1 \leq_p L_2$, then L_2 is* **NP***-complete.*

(c) If L is **NP***-complete and $L \in$* **P***, then* **P**=**NP**.

Notice that parts (a) and (b) have close analogs in computability theory. The analog of part (c) is simply that if L is c.e.-complete then L is undecidable. Part (b) is the basic method for showing new problems are **NP**-complete, and part (c) explains why it is probably a waste of time looking for a polynomial-time algorithm for an **NP**-complete problem.

In practice, a member of **NP** is expressed as a decision problem, and the corresponding language is understood to mean the set of strings coding YES instances to the decision problem using standard coding methods. Thus the problem Satisfiability is: Given a propositional formula F, determine whether F is satisfiable. To show that this is in **NP**, we define the polynomial-time checking relation $R(x, y)$, which holds iff x codes a propositional formula F and y codes a truth assignment to the variables of F that makes F true. This problem was shown in [**9**] to be **NP**-complete essentially by showing that, for each polynomial-time Turing machine M that recognizes a checking relation $R(x, y)$ for an **NP** language L, there is a polynomial-time algorithm that takes as input a string x and produces a propositional formula F_x (describing the computation of M on input (x, y), with variables representing the unknown string y) such that F_x is satisfiable iff M accepts the input (x, y) for some y with $|y| \leq |x|^{O(1)}$.

An important special case of Satisfiability is 3-SAT, which was also shown to be **NP**-complete in [**9**]. Instances of 3-SAT are restricted to formulas in conjunctive normal form with three literals per clause. For example, the formula

$$(P \vee Q \vee R) \wedge (\bar{P} \vee Q \vee \bar{R}) \wedge (P \vee \bar{Q} \vee S) \wedge (\bar{P} \vee \bar{R} \vee \bar{S}) \tag{2}$$

is a YES instance to 3-SAT since the truth assignment τ satisfies the formula, where $\tau(P) = \tau(Q) = True$ and $\tau(R) = \tau(S) = False$.

Many hundreds of **NP**-complete problems have been identified, including SubsetSum (given a set of positive integers presented in decimal notation, and a target T, is there a subset summing to T?), many graph problems (given a graph G, does G have a Hamiltonian cycle? Does G have a clique consisting of half of the vertices? Can the vertices of G be colored with three colors with distinct colors for adjacent vertices?). These problems give rise to many scheduling and routing problems with industrial importance. The book [**15**] provides an excellent reference to the subject, with 300 **NP**-complete problems listed in the appendix.

Associated with each decision problem in **NP** there is a search problem, which is, given a string x, find a string y satisfying the checking relation

$R(x, y)$ for the problem (or determine that x is a NO instance to the problem). Such a y is said to be a *certificate* for x. In the case of an **NP**-complete problem it is easy to see that the search problem can be efficiently reduced to the corresponding decision problem. In fact, if $\mathbf{P} = \mathbf{NP}$, then the associated search problem for every **NP** problem has a polynomial-time algorithm. For example, an algorithm for the decision problem Satisfiability can be used to find a truth assignment τ satisfying a given satisfiable formula F by, for each variable P in F in turn, setting P to True in F or False in F, whichever case keeps F satisfiable.

The set of complements of **NP** languages is denoted co**NP**. The complement of an **NP**-complete language is thought not to be in **NP**; otherwise $\mathbf{NP} = \text{co}\mathbf{NP}$. The set TAUT of tautologies (propositional formulas true under all assignments) is the standard example of a co**NP**-complete language. The conjecture $\mathbf{NP} \neq \text{co}\mathbf{NP}$ is equivalent to the assertion that no formal proof system (suitably defined) for tautologies has short (polynomial-bounded) proofs for all tautologies [**12**]. This fact has motivated the development of a rich theory of propositional proof complexity [**22**], one of whose goals is to prove superpolynomial lower bounds on the lengths of proofs for standard propositional proof systems.

There are interesting examples of **NP** problems not known to be either in **P** or **NP**-complete. One example is the graph isomorphism problem: Given two undirected graphs, determine whether they are isomorphic.

Another example, until recently, was the set of composite numbers. As mentioned in the first section, this set is in **NP**, with checking relation (1), but it is now known also to be in **P** [**1**]. However, the search problem associated with the checking relation (1) is equivalent to the problem of integer factoring and is thought unlikely to be solvable in polynomial time. In fact, an efficient factoring algorithm would break the RSA public key encryption scheme [**28**] commonly used to allow (presumably) secure financial transactions over the Internet.

There is an **NP** decision problem with complexity equivalent to that of integer factoring, namely

$$L_{\text{fact}} = \{\langle a, b\rangle \mid \exists d(1 < d < a \text{ and } d|b)\}.$$

Given an integer $b > 1$, the smallest prime divisor of b can be found with about $\log_2 b$ queries to L_{fact}, using binary search. It is easy to see that the complement of L_{fact} is also in **NP**: a certificate showing $\langle a, b\rangle$ is not in L_{fact} could be the complete prime decomposition of b. Thus L_{fact} is a good example of a problem in **NP** that seems unlikely to be either in **P** or **NP**-complete.

Computational complexity theory plays an important role in modern cryptography [**16**]. The security of the Internet, including most financial

transactions, depends on complexity-theoretic assumptions such as the difficulty of integer factoring or of breaking DES (the Data Encryption Standard). If $\mathbf{P} = \mathbf{NP}$, these assumptions are all false. Specifically, an algorithm solving 3-SAT in n^2 steps could be used to factor 200-digit numbers in a few minutes.

Although a practical algorithm for solving an **NP**-complete problem (showing $\mathbf{P} = \mathbf{NP}$) would have devastating consequences for cryptography, it would also have stunning practical consequences of a more positive nature, and not just because of the efficient solutions to the many **NP**-hard problems important to industry. For example, it would transform mathematics by allowing a computer to find a formal proof of any theorem that has a proof of reasonable length, since formal proofs can easily be recognized in polynomial time. Such theorems may well include all of the CMI prize problems. Although the formal proofs may not be initially intelligible to humans, the problem of finding intelligible proofs would be reduced to that of finding a recognition algorithm for intelligible proofs. Similar remarks apply to diverse creative human endeavors, such as designing airplane wings, creating physical theories, or even composing music. The question in each case is to what extent an efficient algorithm for recognizing a good result can be found. This is a fundamental problem in artificial intelligence, and one whose solution itself would be aided by the **NP**-solver by allowing easy testing of recognition theories.

Even if $\mathbf{P} \neq \mathbf{NP}$ it may still happen that every **NP** problem is susceptible to a polynomial-time algorithm that works on "most" inputs. This could render cryptography impossible and bring about most of the benefits of a world in which $\mathbf{P} = \mathbf{NP}$. This also motivates Levin's theory [**24**], [**18**] of average case completeness, in which the $\mathbf{P} = \mathbf{NP}$ question is replaced by the question of whether every **NP** problem with any reasonable probability distribution on its inputs can be solved in polynomial time on average.

In [**34**] Smale lists the **P** vs **NP** question as problem 3 of mathematical problems for the next century. However, Smale is interested not only in the classical version of the question, but also in a version expressed in terms of the field of complex numbers. Here Turing machines must be replaced by a machine model that is capable of doing exact arithmetic and zero tests on arbitrary complex numbers. The **P** vs **NP** question is replaced by a question related to Hilbert's Nullstellensatz: Is there a polynomial-time algorithm that, given a set of k multivariate polynomials over $\mathbb{C}$, determines whether they have a common zero? See [**4**] for a development of complexity theory in this setting.

The books by Papadimitriou [**25**] and Sipser [**33**] provide good introductions to mainstream complexity theory.

3. The Conjecture and Attempts to Prove It

Most complexity theorists believe that $\mathbf{P} \neq \mathbf{NP}$. Perhaps this can be partly explained by the potentially stunning consequences of $\mathbf{P} = \mathbf{NP}$ mentioned above, but there are better reasons. We explain these by considering the two possibilities in turn: $\mathbf{P} = \mathbf{NP}$ and $\mathbf{P} \neq \mathbf{NP}$.

Suppose first that $\mathbf{P} = \mathbf{NP}$ and consider how one might prove it. The obvious way is to exhibit a polynomial-time algorithm for 3-SAT or one of the other thousand or so known **NP**-complete problems, and, indeed, many false proofs have been presented in this form. There is a standard toolkit available [**7**] for devising polynomial-time algorithms, including the greedy method, dynamic programming, reduction to linear programming, etc. These are the subjects of a course on algorithms, typical in undergraduate computer science curriculums. Because of their importance in industry, a vast number of programmers and engineers have attempted to find efficient algorithms for **NP**-complete problems over the past 30 years, without success. There is a similar strong motivation for breaking the cryptographic schemes that assume $\mathbf{P} \neq \mathbf{NP}$ for their security.

Of course, it is possible that some nonconstructive argument, such as the Robertson–Seymour proofs mentioned earlier in conjunction with the Feasibility Thesis, could show that $\mathbf{P} = \mathbf{NP}$ without yielding any feasible algorithm for the standard **NP**-complete problems. At present, however, the best proven upper bound on an algorithm for solving 3-SAT is approximately 1.5^n, where n is the number of variables in the input formula.

Suppose, on the other hand, that $\mathbf{P} \neq \mathbf{NP}$, and consider how one might prove it. There are two general methods that have been tried: diagonalization with reduction and Boolean circuit lower bounds.

The method of diagonalization with reduction has been used very successfully in computability theory to prove a host of problems undecidable, beginning with the Halting Problem. It has also been used successfully in complexity theory to prove super-exponential lower bounds for very hard decidable problems. For example, Presburger arithmetic, the first-order theory of integers under addition, is a decidable theory for which Fischer and Rabin [**14**] proved that any Turing machine deciding the theory must use at least $2^{2^{cn}}$ steps in the worst case, for some $c > 0$. In general, lower bounds using diagonalization and reduction relativize; that is, they continue to apply in a setting in which both the problem instance and the solving Turing machine can make membership queries to an arbitrary oracle set A. However, in [**3**] it was shown that there is an oracle set A relative to which $\mathbf{P} = \mathbf{NP}$, suggesting that diagonalization with reduction cannot be used to separate these two classes. (There are nonrelativizing results in complexity theory, as

will be mentioned below.) It is interesting to note that relative to a *generic* oracle, $\mathbf{P} \neq \mathbf{NP}$ [**5, 11**].

A *Boolean circuit* is a finite acyclic graph in which each non-input node, or *gate*, is labelled with a Boolean connective; typically from {AND , OR, NOT}. The input nodes are labeled with variables $x_1, ..., x_n$, and for each assignment of 0 or 1 to each variable, the circuit computes a bit value at each gate, including the output gate, in the obvious way. It is not hard to see that if L is a language over $\{0, 1\}$ that is in **P**, then there is a polynomial-size family of Boolean circuits $\langle B_n \rangle$ such that B_n has n inputs, and for each bit string w of length n, when w is applied to the n input nodes of B_n, then the output bit of B_n is 1 iff $w \in L$. In this case we say that $\langle B_n \rangle$ *computes* L.

Thus to prove $\mathbf{P} \neq \mathbf{NP}$ it suffices to prove a super-polynomial lower bound on the size of any family of Boolean circuits solving some specific **NP**-complete problem, such as 3-SAT. Back in 1949 Shannon [**31**] proved that for almost all Boolean functions $f : \{0, 1\}^n \to \{0, 1\}$, any Boolean circuit computing f requires at least $2^n/n$ gates. Unfortunately, his counting argument gives no clue as to how to prove lower bounds for problems in **NP**. Exponential lower bounds for **NP** problems have been proved for restricted circuit models, including monotone circuits [**26**], [**2**] and bounded depth circuits with unbounded fan-in gates [**17**], [**35**] (see [**6**]). However, all attempts to find even super-linear lower bounds for unrestricted Boolean circuits for "explicitly given" Boolean functions have met with total failure; the best such lower bound proved so far is about $4n$. Razborov and Rudich [**27**] explain this failure by pointing out that all methods used so far can be classified as "natural proofs", and natural proofs for general circuit lower bounds are doomed to failure, assuming a certain complexity-theoretic conjecture asserting that strong pseudo-random number generators exist. Since such generators have been constructed assuming the hardness of integer factorization, we can infer the surprising result that a natural proof for a general circuit lower bound would give rise to a more efficient factoring algorithm than is currently known.

The failure of complexity theory to prove interesting lower bounds on a general model of computation is much more pervasive than the failure to prove $\mathbf{P} \neq \mathbf{NP}$. It is consistent with present knowledge that not only could Satisfiability have a polynomial-time algorithm, it could have a linear time algorithm on a multitape Turing machine. The same applies for all 21 problems mentioned in Karp's original paper [**21**]. There are complexity class separations that we know exist but cannot prove. For example, consider the sequence of complexity class inclusions

$$\mathbf{LOGSPACE} \subseteq \mathbf{P} \subseteq \mathbf{NP} \subseteq \mathbf{PSPACE} .$$

A simple diagonal argument shows that the first is a proper subset of the last, but we cannot prove any particular adjacent inclusion is proper.

As another example, let **LINEAR-SIZE** be the class of languages over $\{0,1\}$ that can be computed by a family $\langle B_n \rangle$ of Boolean circuits of size $O(n)$. It is not known whether either **P** or **NP** is a subset of **LINEAR-SIZE**, although Kannan [**20**] proved that there are languages in the polynomial hierarchy (a generalization of **NP**) not in **LINEAR-SIZE**. Since if $\mathbf{P} = \mathbf{NP}$, the polynomial hierarchy collapses to **P**, we have

PROPOSITION 2. *If* $\mathbf{P} \subseteq$ **LINEAR-SIZE**, *then* $\mathbf{P} \neq \mathbf{NP}$.

This proposition could be interpreted as a method of proving $\mathbf{P} \neq \mathbf{NP}$, but a more usual belief is that the hypothesis is false.

A fundamental question in complexity theory is whether a source of random bits can be used to speed up substantially the recognition of some languages, provided one is willing to accept a small probability of error. The class **BPP** consists of all languages L that can be recognized by a randomized polynomial-time algorithm, with at most an exponentially small error probability on every input. Of course $\mathbf{P} \subseteq \mathbf{BPP}$, but it is not known whether the inclusion is proper. The set of prime numbers is in **BPP** [**36**], but it is not known to be in **P**. A reason for thinking that $\mathbf{BPP} = \mathbf{P}$ is that randomized algorithms are often successfully executed using a deterministic pseudo-random number generator as a source of "random" bits.

There is an indirect but intriguing connection between the two questions $\mathbf{P} = \mathbf{BPP}$ and $\mathbf{P} = \mathbf{NP}$. Let **E** be the class of languages recognizable in exponential time; that is the class of languages L such that $L = L(M)$ for some Turing machine M with $T_M(n) = O(2^{cn})$ for some $c > 0$. Let **A** be the assertion that some language in **E** requires exponential circuit complexity. That is,

ASSERTION A. *There is* $L \in \mathbf{E}$ *and* $\epsilon > 0$ *such that, for every circuit family* $\langle B_n \rangle$ *computing* L *and for all sufficiently large* n, B_n *has at least* $2^{\epsilon n}$ *gates.*

PROPOSITION 3. *If* **A** *then* $\mathbf{BPP} = \mathbf{P}$. *If* **not A** *then* $\mathbf{P} \neq \mathbf{NP}$.

The first implication is a lovely theorem by Impagliazzo and Wigderson [**19**] and the second is an intriguing observation by V. Kabanets that strengthens Proposition 2. In fact, Kabanets concludes $\mathbf{P} \neq \mathbf{NP}$ from a weaker assumption than **not A**; namely that every language in **E** can be computed by a Boolean circuit family $\langle B_n \rangle$ such that for at least one n, B_n has fewer gates than the maximum needed to compute any Boolean function $f : \{0,1\}^n \to \{0,1\}$. But there is no consensus on whether this hypothesis is true.

We should point out that Proposition 3 relativizes, and, in particular, relative to any **PSPACE**-complete oracle Assertion A holds and $\mathbf{BPP} = \mathbf{P} = \mathbf{NP}$. Thus a nonrelativizing construction will be needed if one is to

prove $\mathbf{P} \neq \mathbf{NP}$ by giving small circuits for languages in **E**. Nonrelativizing constructions have been used successfully before, for example in showing **IP** (Interactive Polynomial time) contains all of **PSPSACE** [**30**]. In this and other such constructions a key technique is to represent Boolean functions by multivariate polynomials over finite fields.

Appendix: Definition of Turing Machine

A Turing machine M consists of a finite state control (i.e., a finite program) attached to a read/write head moving on an infinite tape. The tape is divided into squares, each capable of storing one symbol from a finite alphabet Γ that includes the blank symbol b. Each machine M has a specified input alphabet Σ, which is a subset of Γ, not including the blank symbol b. At each step in a computation, M is in some state q in a specified finite set Q of possible states. Initially, a finite input string over Σ is written on adjacent squares of the tape, all other squares are blank (contain b), the head scans the left-most symbol of the input string, and M is in the initial state q_0. At each step M is in some state q and the head is scanning a tape square containing some tape symbol s, and the action performed depends on the pair (q, s) and is specified by the machine's transition function (or program) δ. The action consists of printing a symbol on the scanned square, moving the head left or right one square, and assuming a new state.

Formally, a Turing machine M is a tuple $\langle \Sigma, \Gamma, Q, \delta \rangle$, where Σ, Γ, Q are finite nonempty sets with $\Sigma \subseteq \Gamma$ and $b \in \Gamma - \Sigma$. The state set Q contains three special states $q_0, q_{\text{accept}}, q_{\text{reject}}$. The *transition function* δ satisfies

$$\delta : (Q - \{q_{\text{accept}}, q_{\text{reject}}\}) \times \Gamma \to Q \times \Gamma \times \{-1, 1\}.$$

If $\delta(q, s) = (q', s', h)$, the interpretation is that, if M is in state q scanning the symbol s, then q' is the new state, s' is the symbol printed, and the tape head moves left or right one square depending on whether h is -1 or 1.

We assume that the sets Q and Γ are disjoint.

A *configuration* of M is a string xqy with $x, y \in \Gamma^*$, y not the empty string, and $q \in Q$.

The interpretation of the configuration xqy is that M is in state q with xy on its tape, with its head scanning the left-most symbol of y.

If C and C' are configurations, then $C \xrightarrow{M} C'$ if $C = xqsy$ and $\delta(q, s) = (q', s', h)$ and one of the following holds:

$C' = xs'q'y$ and $h = 1$ and y is nonempty.

$C' = xs'q'b$ and $h = 1$ and y is empty.

$C' = x'q'as'y$ and $h = -1$ and $x = x'a$ for some $a \in \Gamma$.

$C' = q'bs'y$ and $h = -1$ and x is empty.

A configuration xqy is *halting* if $q \in \{q_{\text{accept}}, q_{\text{reject}}\}$. Note that for each non-halting configuration C there is a unique configuration C' such that $C \xrightarrow{M} C'$.

The *computation* of M on input $w \in \Sigma^*$ is the unique sequence $C_0, C_1, ...$ of configurations such that $C_0 = q_0 w$ (or $C_0 = q_0 b$ if w is empty) and $C_i \xrightarrow{M} C_{i+1}$ for each i with C_{i+1} in the computation, and either the sequence is infinite or it ends in a halting configuration. If the computation is finite, then the number of steps is one less than the number of configurations; otherwise the number of steps is infinite. We say that M *accepts* w iff the computation is finite and the final configuration contains the state q_{accept}.

Acknowledgments

My thanks to Avi Wigderson and Hugh Woodin for many helpful suggestions for improving an earlier version of this paper.

Bibliography

[1] M. Agrawal, N. Kayal, and N. Saxena, *Primes is in P*, Ann. Math. **160** (2004), 781–793.

[2] N. Alon and R.B. Boppana, *The monotone circuit complexity of boolean functions*, Combinatorica **7** (1987), 1–22.

[3] T. Baker, J. Gill, and R. Solovay, *Relativizations of the P =? NP question*, SICOMP: SIAM Journal on Computing, 1975.

[4] L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity and Real Computation*, Springer-Verlag, New York, 1998.

[5] M. Blum and R. Impagliazzo, *Generic oracles and oracle classes*, in Proceedings of the 28th Annual Symposium on Foundations of Computer Science, A.K. Chandra, ed., IEEE Computer Society Press, Los Angeles, 1987, 118–126.

[6] R.B. Boppana and M. Sipser, *The complexity of finite functions*, Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity, J. Van Leeuwen, ed., Elsevier and The MIT Press, Cambridge, MA, 1990, 759–804.

[7] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd edition, McGraw Hill, New York, 2001.

[8] A. Cobham, *The intrinsic computational difficulty of functions*, in Proceedings of the 1964 International Congress for Logic, Methodology, and Philosophy of Science, Y. Bar-Hille, ed., Elsevier/North-Holland, Amsterdam, 1964, 24–30.

[9] S. Cook, *The complexity of theorem-proving procedures*, in Conference Record of Third Annual ACM Symposium on Theory of Computing, ACM, New York, 1971, 151–158.

[10] S. Cook, *Computational complexity of higher type functions*, in Proceedings of the International Congress of Mathematicians, Kyoto, Japan, Springer-Verlag, Berlin, 1991, 55–69.

[11] S. Cook, R. Impagliazzo, and T. Yamakami, *A tight relationship between generic oracles and type-2 complexity theory*, Information and Computation **137** (1997), 159–170.

[12] S. Cook and R. Reckhow, *The relative efficiency of propositional proof systems*, J. Symbolic Logic **44** (1979), 36–50.

[13] J. Edmonds, *Minimum partition of a matroid into independent subsets*, J. Res. Nat. Bur. Standards Sect. B **69** (1965), 67–72.

[14] M.J. Fischer and M.O. Rabin, *Super-exponential complexity of Presburger arithmetic*, in Complexity of Computation **7**, AMS, Providence, RI, 1974, 27–41.
[15] M.R. Garey and D.S. Johnson, *Computers and Intractability, a Guide to the Theory of NP-Completeness*, W.H. Freeman and Co., San Francisco, 1979.
[16] O. Goldreich, *The Foundations of Cryptography — Volume 1*, Cambridge University Press, Cambridge, UK, 2000.
[17] J. Hastad, *Almost optimal lower bounds for small depth circuits*, in Randomness and Computation, Advances in Computing Research **5**, JAI Press Inc., Greenwich, CT, 1989, 143–170.
[18] R. Impagliazzo, *A personal view of average-case complexity*, in 10th IEEE Annual Conference on Structure in Complexity Theory, IEEE Computer Society Press, Washington, DC, 1995, 134–147.
[19] R. Impagliazzo and A. Wigderson, P = BPP *if* E *requires exponential circuits: Derandomizing the* XOR *lemma*, in ACM Symposium on Theory of Computing (STOC), ACM, New York, 1997, 220–229.
[20] R. Kannan, *Circuit-size lower bounds and non-reducibility to sparse sets*, Information and Control **55** (1982), 40–56..
[21] R.M. Karp, *Reducibility among combinatorial problems*, in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, 85–103.
[22] J. Krajicek, *Bounded Arithmetic, Propositional Logic, and Complexity Theory*, Cambridge University Press, Cambridge, 1995.
[23] L. Levin, *Universal search problems* (in Russian), Problemy Peredachi Informatsii *9* (1973), 265–266. English translation in B. A. Trakhtenbrot, *A survey of Russian approaches to Perebor (brute-force search) algorithms*, Annals of the History of Computing **6** (1984), 384-400.
[24] L. Levin, *Average case complete problems*, SIAM J. Computing **15** (1986), 285–286.
[25] C. Papadimitriou, *Computational Complexity*, Addison-Wesley, Reading, MA, 1994.
[26] A.A. Razborov, *Lower bounds on the monotone complexity of some boolean functions*, Soviet Math. Dokl. **31** (1985), 354–357.
[27] A.A. Razborov and S. Rudich, *Natural proofs*, Journal of Computer and System Sciences **55** (1997), 24–35.
[28] R.L. Rivest, A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Comm. ACM **21** (1978), 120–126.
[29] N. Robertson and P.D. Seymour, *Graph minors i-xiii*, Journal of Combinatorial Theory B, 1983–1995.
[30] A. Shamir, IP = PSPACE, J.A.C.M. **39** (1992), 869–977.
[31] C. Shannon, *The synthesis of two-terminal switching circuits*, Bell System Technical Journal **28** (1949), 59–98.
[32] P. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM J. Computing **26** (1997), 1484–1509.
[33] M. Sipser, *Introduction to the Theory of Computation*, PWS Publ., Boston, 1997.
[34] S. Smale, *Mathematical problems for the next century*, Math. Intelligencer **20**, no. 2, 1998, 7–15.
[35] R. Smolensky, *Algebraic methods in the theory of lower bounds for boolean circuit complexity*, in ACM Symposium on Theory of Computing (STOC) **19**, ACM, New York, 1987, 77-82.
[36] R. Solovay and V. Strassen, *A fast Monte-Carlo test for primality*, SIAM Journal on Computing **6** (1977), 84–85.
[37] A. Turing, *On computable numbers with an application to the entscheidnungsproblem*, Proc. London Math. Soc. **42** (1936), 230–265.

[38] J. von Neumann, *A certain zero-sum two-person game equivalent to the optimal assignment problem*, in Contributions to the Theory of Games II, H.W. Kahn and A.W. Tucker, eds. Princeton Univ. Press, Princeton, NJ, 1953, 5–12.

The Riemann Hypothesis

Enrico Bombieri

The Riemann Hypothesis

Enrico Bombieri

1. The Problem

The Riemann zeta function is the function of the complex variable s, defined in the half-plane[1] $\Re(s) > 1$ by the absolutely convergent series

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s},$$

and in the whole complex plane $\mathbb{C}$ by analytic continuation. As shown by Riemann, $\zeta(s)$ extends to $\mathbb{C}$ as a meromorphic function with only a simple pole at $s = 1$, with residue 1, and satisfies the functional equation

$$\pi^{-s/2}\Gamma\left(\frac{s}{2}\right)\zeta(s) = \pi^{-(1-s)/2}\Gamma\left(\frac{1-s}{2}\right)\zeta(1-s). \tag{1}$$

In an epoch-making memoir published in 1859, Riemann [**18**] obtained an analytic formula for the number of primes up to a preassigned limit. This formula is expressed in terms of the zeros of the zeta function, namely the solutions $\rho \in \mathbb{C}$ of the equation $\zeta(\rho) = 0$.

In this paper, Riemann introduces the function of the complex variable t defined by

$$\xi(t) = \frac{1}{2}s(s-1)\pi^{-s/2}\Gamma\left(\frac{s}{2}\right)\zeta(s),$$

with $s = \frac{1}{2} + it$, and shows that $\xi(t)$ is an even entire function of t whose zeros have imaginary part between $-i/2$ and $i/2$. He further states, sketching a proof, that in the range between 0 and T the function $\xi(t)$ has about $(T/2\pi)\log(T/2\pi) - T/2\pi$ zeros. Riemann then continues "Man findet nun in der That etwa so viel reelle Wurzeln innerhalb dieser Grenzen, und es ist sehr wahrscheinlich, dass alle Wurzeln reell sind," which can be translated as "Indeed, one finds between those limits about that many real zeros, and it is very likely that all zeros are real."

[1]We denote by $\Re(s)$ and $\Im(s)$ the real and imaginary part of the complex variable s. The use of the variable s is already in Dirichlet's famous work of 1837 on primes in arithmetic progression.

The statement that all zeros of the function $\xi(t)$ are real is the Riemann hypothesis.

The function $\zeta(s)$ has zeros at the negative even integers $-2, -4, \ldots$ and one refers to them as the *trivial zeros*. The other zeros are the complex numbers $\frac{1}{2} + i\alpha$, where α is a zero of $\xi(t)$. Thus, in terms of the function $\zeta(s)$, we can state the

RIEMANN HYPOTHESIS. *The nontrivial zeros of* $\zeta(s)$ *have real part equal to* $\frac{1}{2}$.

In the opinion of many mathematicians, the Riemann hypothesis, and its extension to general classes of L-functions, is probably the most important open problem in pure mathematics today.

2. History and Significance of the Riemann Hypothesis

For references pertaining to the early history of zeta functions and the theory of prime numbers, we refer to Landau [**13**] and Edwards [**6**].

The connection between prime numbers and the zeta function, by means of the celebrated *Euler product*

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

valid for $\Re(s) > 1$, appears for the first time in Euler's book *Introductio in Analysin Infinitorum*, published in 1748. Euler also studied the values of $\zeta(s)$ at the even positive and negative integers, and he divined a functional equation, equivalent to Riemann's functional equation, for the closely related function $\sum(-1)^{n-1}/n^s$ (see the interesting account of Euler's work in Hardy's book [**8**]).

The problem of the distribution of prime numbers received attention for the first time with Gauss and Legendre, at the end of the eighteenth century. Gauss, in a letter to the astronomer Hencke in 1849, stated that he had found in his early years that the number $\pi(x)$ of primes up to x is well approximated by the function[2]

$$\mathrm{Li}(x) = \int_0^x \frac{dt}{\log t}.$$

In 1837, Dirichlet proved his famous theorem of the existence of infinitely many primes in any arithmetic progression $qn + a$ with q and a positive coprime integers.

On May 24, 1848, Tchebychev read at the Academy of St. Petersburg his first memoir on the distribution of prime numbers, later published in

[2]The integral is a principal value in the sense of Cauchy.

1850. It contains the first study of the function $\pi(x)$ by analytic methods. Tchebychev begins by taking the logarithm of the Euler product, obtaining[3]

$$-\sum_p \log\left(1-\frac{1}{p^s}\right)+\log(s-1)=\log\left((s-1)\zeta(s)\right), \tag{2}$$

which is his starting point.

Next, he proves the integral formula

$$\zeta(s)-1-\frac{1}{s-1}=\frac{1}{\Gamma(s)}\int_0^\infty\left(\frac{1}{e^x-1}-\frac{1}{x}\right)e^{-x}x^{s-1}dx, \tag{3}$$

out of which he deduces that $(s-1)\zeta(s)$ has limit 1, and also has finite derivatives of any order, as s tends to 1 from the right. He then observes that the derivatives of any order of the left-hand side of (2) can be written as a fraction in which the numerator is a polynomial in the derivatives of $(s-1)\zeta(s)$, and the denominator is an integral power of $(s-1)\zeta(s)$, from which it follows that the right-hand side of (2) has finite derivatives of any order, as s tends to 1 from the right. From this, he is able to prove that if there is an asymptotic formula for $\pi(x)$ by means of a finite sum $\sum a_k x/(\log x)^k$, up to an order $O(x/(\log x)^N)$, then $a_k=(k-1)!$ for $k=1,\dots,N-1$. This is precisely the asymptotic expansion of the function $\mathrm{Li}(x)$, thus vindicating Gauss's intuition.

A second paper by Tchebychev gave rigorous proofs of explicit upper and lower bounds for $\pi(x)$, of the correct order of magnitude. Here, he introduces the counting functions

$$\vartheta(x)=\sum_{p\le x}\log p,\qquad \psi(x)=\vartheta(x)+\vartheta(\sqrt[2]{x})+\vartheta(\sqrt[3]{x})+\cdots$$

and proves the identity[4]

$$\sum_{n\le x}\psi\left(\frac{x}{n}\right)=\log[x]!\,.$$

From this identity, he finally obtains numerical upper and lower bounds for $\psi(x)$, $\vartheta(x)$ and $\pi(x)$.

Popular variants of Tchebychev's method, based on the integrality of suitable ratios of factorials, originate much later and cannot be ascribed to Tchebychev.

Riemann's memoir on $\pi(x)$ is really astonishing for the novelty of ideas introduced. He first writes $\zeta(s)$ using the integral formula, valid for $\Re(s)>1$:

$$\zeta(s)=\frac{1}{\Gamma(s)}\int_0^\infty\frac{e^{-x}}{1-e^{-x}}x^{s-1}dx, \tag{4}$$

[3]Tchebychev uses $1+\rho$ in place of our s. We write his formulas in modern notation.

[4]Here $[x]$ denotes the integral part of x.

and then deforms the contour of integration in the complex plane, so as to obtain a representation valid for any s. This gives the analytic continuation and the functional equation of $\zeta(s)$. Then he gives a second proof of the functional equation in the symmetric form (1), introduces the function $\xi(t)$ and states some of its properties as a function of the complex variable t.

Riemann continues by writing the logarithm of the Euler product as an integral transform, valid for $\Re(s) > 1$:

$$\frac{1}{s}\log\zeta(s) = \int_1^\infty \Pi(x)x^{-s-1}dx \tag{5}$$

where

$$\Pi(x) = \pi(x) + \frac{1}{2}\pi(\sqrt[2]{x}) + \frac{1}{3}\pi(\sqrt[3]{x}) + \cdots .$$

By Fourier inversion, he is able to express $\Pi(x)$ as a complex integral, and compute it using the calculus of residues. The residues occur at the singularities of $\log\zeta(s)$ at $s = 1$ and at the zeros of $\zeta(s)$. Finally an inversion formula expressing $\pi(x)$ in terms of $\Pi(x)$ yields Riemann's formula.

This was a remarkable achievement that immediately attracted much attention. Even if Riemann's initial line of attack may have been influenced by Tchebychev (we find several explicit references to Tchebychev in Riemann's unpublished Nachlass[5]), his great contribution was to see how the distribution of prime numbers is determined by the complex zeros of the zeta function.

At first sight, the Riemann hypothesis appears to be only a plausible interesting property of the special function $\zeta(s)$, and Riemann himself seems to take that view. He writes: "Hiervon wäre allerdings ein strenger Beweis zu wünschen; ich habe indess die Aufsuchung desselben nach einigen flüchtigen vergeblichen Versuchen vorläufig bei Seite gelassen, da er für den nächsten Zweck meiner Untersuchung entbehrlich schien," which can be translated as "Without doubt it would be desirable to have a rigorous proof of this proposition; however I have left this research aside for the time being after some quick unsuccessful attempts, because it appears to be unnecessary for the immediate goal of my study."

On the other hand, one should not draw from this comment the conclusion that the Riemann hypothesis was only a casual remark of minor interest for him. The validity of the Riemann hypothesis is equivalent to saying that the deviation of the number of primes from the mean $\mathrm{Li}(x)$ is

$$\pi(x) = \mathrm{Li}(x) + O\left(\sqrt{x}\log x\right);$$

[5]The Nachlass consists of Riemann's unpublished notes and is preserved in the mathematical library of the University of Göttingen. The part regarding the zeta function was analyzed in depth by C.L. Siegel [**22**].

the error term cannot be improved by much, since it is known to oscillate in both directions to order at least $\mathrm{Li}(\sqrt{x})\log\log\log x$ (Littlewood). In view of Riemann's comments at the end of his memoir about the approximation of $\pi(x)$ by $\mathrm{Li}(x)$, it is quite likely that he saw how his hypothesis was central to the question of how good an approximation to $\pi(x)$ one may get from his formula.

The failure of the Riemann hypothesis would create havoc in the distribution of prime numbers. This fact alone singles out the Riemann hypothesis as the main open question of prime number theory.

The Riemann hypothesis has become a central problem of pure mathematics, and not just because of its fundamental consequences for the law of distribution of prime numbers. One reason is that the Riemann zeta function is not an isolated object, rather it is the prototype of a general class of functions, called *L-functions*, associated with algebraic (automorphic representations) or arithmetical objects (arithmetic varieties); we shall refer to them as *global L-functions*. They are Dirichlet series with a suitable Euler product and are expected to satisfy an appropriate functional equation and a Riemann hypothesis. The factors of the Euler product may also be considered as some kind of zeta functions of a local nature, which also should satisfy an appropriate Riemann hypothesis (the so-called Ramanujan property). The most important properties of the algebraic or arithmetical objects underlying an L-function can or should be described in terms of the location of its zeros and poles, and values at special points.

The consequences of a Riemann hypothesis for global L-functions are important and varied. We mention here, to indicate the variety of situations to which it can be applied, an extremely strong effective form of Tchebotarev's density theorem for number fields, the non-trivial representability of 0 by a non-singular cubic form in seven or more variables (provided it satisfies the appropriate necessary congruence conditions for solubility, (Hooley, [**9**])), and Miller's deterministic polynomial time primality test. On the other hand, many deep results in number theory that are consequences of a general Riemann hypothesis can be shown to hold independent of it, thus adding considerable weight to the validity of the conjecture.

It is outside the scope of this article even to outline the definition of global L-functions, referring instead to Iwaniec and Sarnak [**10**] for a survey of the expected properties satisfied by them; it suffices here to say that the study of the analytic properties of these functions presents extraordinary difficulties.

Already the analytic continuation of L-functions as meromorphic or entire functions is known only in special cases. For example, the functional equation for the L-function of an elliptic curve over $\mathbb{Q}$ and for its twists by

Dirichlet characters is an easy consequence of, and is equivalent to, the existence of a parametrization of the curve by means of modular functions for a Hecke group $\Gamma_0(N)$; the real difficulty lies in establishing this modularity. No one knows how to prove this functional equation by analytic methods, but the modularity of elliptic curves over $\mathbb{Q}$ has been established directly, first in the semistable case in the spectacular work of Wiles [**28**] and Taylor and Wiles [**24**] leading to the solution of Fermat's Last Theorem, and then in the general case in a recent preprint by Breuil, Conrad, Diamond and Taylor.

Not all L-functions are directly associated to arithmetic or geometric objects. The simplest example of L-functions not of arithmetic or geometric nature are those arising from Maass waveforms for a Riemann surface X uniformized by an arithmetic subgroup Γ of $\mathrm{PGL}(2,\mathbb{R})$. They are pull-backs $f(z)$ to the universal covering space $\Im(z) > 0$ of X, of simultaneous eigenfunctions for the action of the hyperbolic Laplacian and of the Hecke operators on X.

The most important case is again the group $\Gamma_0(N)$. In this case one can introduce a notion of *primitive* waveform, analogous to the notion of primitive Dirichlet character, meaning that the waveform is not induced from another waveform for a $\Gamma_0(N')$ with N' a proper divisor of N. For a primitive waveform, the action of the Hecke operators T_n is defined for every n, and the L-function can be defined as $\sum \lambda_f(n) n^{-s}$, where $\lambda_f(n)$ is the eigenvalue of T_n acting on the waveform $f(z)$. Such an L-function has an Euler product and satisfies a functional equation analogous to that for $\zeta(s)$. It is also expected to satisfy a Riemann hypothesis.

Not a single example of validity or failure of a Riemann hypothesis for an L-function is known up to this date. The Riemann hypothesis for $\zeta(s)$ does not seem to be any easier than for Dirichlet L-functions (except possibly for non-trivial real zeros), leading to the view that its solution may require attacking much more general problems, by means of entirely new ideas.

3. Evidence for the Riemann Hypothesis

Notwithstanding some skepticism voiced in the past, based perhaps more on the number of failed attempts to a proof rather than on solid heuristics, it is fair to say that today there is quite a bit of evidence in its favor. We have already emphasized that the general Riemann hypothesis is consistent with our present knowledge of number theory. There is also specific evidence of a more direct nature, which we shall now examine.

First, strong numerical evidence.

Interestingly enough, the first numerical computation of the first few zeros of the zeta function already appears in Riemann's Nachlass. A rigorous

verification of the Riemann hypothesis in a given range can be done numerically as follows. The number $N(T)$ of zeros of $\zeta(s)$ in the rectangle $\mathcal{R}$ with vertices at $-1-iT, 2-iT, 2+iT, -1+iT$ is given by Cauchy's integral

$$N(T) - 1 = \frac{1}{2\pi i} \int_{\partial \mathcal{R}} -\frac{\zeta'}{\zeta}(s) ds,$$

provided T is not the imaginary part of a zero (the -1 in the left-hand side of this formula is due to the simple pole of $\zeta(s)$ at $s = 1$). The zeta function and its derivative can be computed to arbitrary high precision using the MacLaurin summation formula or the Riemann–Siegel formula [**22**]; the quantity $N(T)-1$, which is an integer, is then computed exactly by dividing by $2\pi i$ the numerical evaluation of the integral, and rounding off its real part to the nearest integer (this is only of theoretical interest, and much better methods are available in practice for computing $N(T)$ exactly). On the other hand, since $\xi(t)$ is continuous and real for real t, there will be a zero of odd order between any two points at which $\xi(t)$ changes sign. By judiciously choosing sample points, one can detect sign changes of $\xi(t)$ in the interval $[-T, T]$. If the number of sign changes equals $N(T)$, one concludes that all zeros of $\zeta(s)$ in $\mathcal{R}$ are simple and satisfy the Riemann hypothesis. In this way, it has been shown by van de Lune, te Riele and Winter [**15**] that the first 1.5 billion zeros of $\zeta(s)$, arranged by increasing positive imaginary part, are simple and satisfy the Riemann hypothesis.

The Riemann hypothesis is equivalent to the statement that all local maxima of $\xi(t)$ are positive and all local minima are negative, and it has been suggested that if a counterexample exists, then it should be in the neighborhood of unusually large peaks of $|\zeta(\frac{1}{2} + it)|$. The above range for T is $T \cong 5 \times 10^8$ and is not large enough for $|\zeta(\frac{1}{2} + it)|$ to exhibit these peaks, which are known to occur eventually. Further calculations done by Odlyzko [**17**] in selected intervals show that the Riemann hypothesis holds for over 3×10^8 zeros at heights up to[6] 2×10^{20}. These calculations also strongly support independent conjectures by Dyson and Montgomery [**16**] concerning the distribution of spacings between zeros.

Computing zeros of L-functions is more difficult, but this has been done in several cases, including examples of Dirichlet L-functions, L-functions of elliptic curves, Maass L-functions and nonabelian Artin L-functions arising from number fields of small degree. No exception to a generalized Riemann hypothesis has been found in this way.

Second, it is known that hypothetical exceptions to the Riemann hypothesis must be rare if we move away from the line $\Re(s) = \frac{1}{2}$.

[6]The most recent calculations by Odlyzko, which are approaching completion, will explore completely the interval $[10^{22}, 10^{22} + 10^{10}]$.

Let $N(\alpha, T)$ be the number of zeros of $\zeta(s)$ in the rectangle $\alpha \leq \Re(s) \leq 2$, $0 \leq \Im(s) \leq T$. The prototype result goes back to Bohr and Landau in 1914, namely $N(\alpha, T) = O(T)$ for any fixed α with $\frac{1}{2} < \alpha < 1$. A significant improvement of the result of Bohr and Landau was obtained by Carlson in 1920, obtaining the *density theorem* $N(\alpha, T) = O(T^{4\alpha(1-\alpha)+\varepsilon})$ for any fixed $\varepsilon > 0$. The fact that the exponent here is strictly less than 1 is important for arithmetic applications, for example, in the study of primes in short intervals. The exponent in Carlson's theorem has gone through several successive refinements for various ranges of α, in particular in the range $\frac{3}{4} < \alpha < 1$. Curiously enough, the best exponent known to date in the range $\frac{1}{2} < \alpha \leq \frac{3}{4}$ remains Ingham's exponent $3(1-\alpha)/(2-\alpha)$, obtained in 1940. For references to these results, the reader may consult the recent revision by Heath-Brown of the classical monograph of Titchmarsh [**23**], and the book by Ivič [**11**].

Third, it is known that more than 40% of nontrivial zeros of $\zeta(s)$ are simple and satisfy the Riemann hypothesis (Selberg [**20**], Levinson [**14**], Conrey [**2**]). Most of these results have been extended to other L-functions, including all Dirichlet L-functions and L-functions associated to modular forms or Maass waveforms.

4. Further Evidence: Varieties Over Finite Fields

It may be said that the best evidence in favor of the Riemann hypothesis derives from the corresponding theory, which has been developed in the context of algebraic varieties over finite fields. The simplest situation is as follows.

Let C be a nonsingular projective curve over a finite field $\mathbb{F}_q$ of characteristic p with $q = p^a$ elements. Let $\mathrm{Div}(C)$ be the additive group of divisors on C defined over $\mathbb{F}_q$, in other words, formal finite sums $\mathfrak{a} = \sum a_i P_i$ with $a_i \in \mathbb{Z}$ and P_i points of C defined over a finite extension of $\mathbb{F}_q$, such that $\phi(\mathfrak{a}) = \mathfrak{a}$ where ϕ is the Frobenius endomorphism on C raising coordinates to the qth power. The quantity $\deg(\mathfrak{a}) = \sum a_i$ is the degree of the divisor $\mathfrak{a}$. The divisor $\mathfrak{a}$ is called effective if every a_i is a positive integer; in this case, we write $\mathfrak{a} > 0$. Finally, a prime divisor $\mathfrak{p}$ is a positive divisor that cannot be expressed as the sum of two positive divisors. By definition, the norm of a divisor $\mathfrak{a}$ is $N\mathfrak{a} = q^{\deg(\mathfrak{a})}$.

The zeta function of the curve C, as defined by E. Artin, H. Hasse and F.K. Schmidt, is

$$\zeta(s, C) = \sum_{\mathfrak{a}>0} \frac{1}{N\mathfrak{a}^s}.$$

This function has an Euler product

$$\zeta(s, C) = \prod_{\mathfrak{p}} (1 - N\mathfrak{p}^{-s})^{-1}$$

and a functional equation

$$q^{(g-1)s}\zeta(s, C) = q^{(g-1)(1-s)}\zeta(1 - s, C),$$

where g is the genus of the curve C; it is a consequence of the Riemann–Roch theorem. The function $\zeta(s, C)$ is a rational function of the variable $t = q^{-s}$, hence is periodic[7] with period $2\pi i/\log q$ and has simple poles at the points $s = 2\pi i m/\log q$ and $s = 1 + 2\pi i m/\log q$ for $m \in \mathbb{Z}$. Expressed in terms of the variable t, the zeta function becomes a rational function $Z(t, C)$ of t, with simple poles at $t = 1$ and $t = q^{-1}$. The use of the variable t, rather than q^{-s}, is more natural in the geometric case and we refer to Zeta functions, with a capital Z, to indicate the corresponding objects.

The Riemann hypothesis for $\zeta(s, C)$ is the statement that all its zeros have real part equal to $\frac{1}{2}$; in terms of the Zeta function $Z(t, C)$, which has a numerator of degree $2g$, has zeros of absolute value $q^{-1/2}$.

This is easy to verify if $g = 0$, because the numerator is 1. For $g = 1$, a proof was obtained by Hasse in 1934. The general case of arbitrary genus g was finally settled by Weil in the early 1940s (see his letter to E. Artin of July 10, 1942, where he gives a complete sketch of the theory of correspondences on a curve [**25**]); his results were eventually published in book form in 1948 [**26**].

Through his researches, Weil was led to the formulation of sweeping conjectures about Zeta functions of general algebraic varieties over finite fields, relating their properties to the topological structure of the underlying algebraic variety. Here the Riemann hypothesis, in a simplified form, is the statement that the reciprocals of the zeros and poles of the Zeta function (the so-called *characteristic roots*) have absolute value $q^{d/2}$ with d a positive integer or 0, and are interpreted as eigenvalues of the Frobenius automorphism acting on the cohomology of the variety. After M. Artin, A. Grothendieck, and J.-L. Verdier developed the fundamental tool of étale cohomology, the proof of the corresponding Riemann hypothesis for Zeta functions of arbitrary varieties over finite fields was finally obtained by Deligne [**3**], [**4**]. Deligne's theorem surely ranks as one of the crowning achievements of 20th century mathematics. Its numerous applications to the solution of long-standing problems in number theory, algebraic geometry, and discrete mathematics are witness to the significance of these general Riemann hypotheses.

[7]Similarly, $\zeta(s)$ is almost periodic in any half-plane $\Re(s) \geq 1 + \delta$, $\delta > 0$.

In our opinion, these results in the geometric setting cannot be ignored as not relevant to the understanding of the classical Riemann hypothesis; the analogies are too compelling to be dismissed outright.

5. Further Evidence: The Explicit Formula

A conceptually important generalization of Riemann's explicit formula for $\pi(x)$, obtained by Weil [**27**] in 1952, offers a clue to what may still lie undiscovered behind the problem.

Consider the class $\mathcal{W}$ of complex-valued functions $f(x)$ on the positive half-line $\mathbb{R}_+$, continuous and continuously differentiable except for finitely many points at which both $f(x)$ and $f'(x)$ have at most a discontinuity of the first kind, and at which the value of $f(x)$ and $f'(x)$ is defined as the average of the right and left limits there. Suppose also that there is $\delta > 0$ such that $f(x) = O(x^{\delta})$ as $x \to 0+$ and $f(x) = O(x^{-1-\delta})$ as $x \to +\infty$.

Let $\widetilde{f}(s)$ be the Mellin transform

$$\widetilde{f}(s) = \int_0^{\infty} f(x) x^s \frac{dx}{x},$$

which is an analytic function of s for $-\delta < \Re(s) < 1 + \delta$.

For the Riemann zeta function, Weil's formula can be stated as follows. Let $\Lambda(n) = \log p$ if $n = p^a$ is a power of a prime p, and 0 otherwise. We have

Explicit Formula. *For $f \in \mathcal{W}$ we have*

$$\widetilde{f}(0) - \sum_{\rho} \widetilde{f}(\rho) + \widetilde{f}(1) = \sum_{n=1}^{\infty} \Lambda(n) \left\{ f(n) + \frac{1}{n} f\left(\frac{1}{n}\right) \right\} + (\log 4\pi + \gamma) f(1) + \int_1^{\infty} \left\{ f(x) + \frac{1}{x} f\left(\frac{1}{x}\right) - \frac{2}{x} f(1) \right\} \frac{dx}{x - x^{-1}}.$$

Here the first sum ranges over all nontrivial zeros of $\zeta(s)$ and is understood as

$$\lim_{T \to +\infty} \sum_{|\Im(\rho)| < T} \widetilde{f}(\rho).$$

In his paper, Weil showed that there is a corresponding formula for zeta and L-functions of number fields as well as for Zeta functions of curves over finite fields. The terms in the right-hand side of the equation can be written as a sum of terms of local nature, associated to the absolute values of the underlying number field, or function field in the case of curves over a field of positive characteristic. Moreover, in the latter case the explicit formula can be deduced from the Lefschetz fixed point formula, applied to the Frobenius endomorphism on the curve C. The three terms in the

left-hand side, namely $\widetilde{f}(0)$, $\sum \widetilde{f}(\rho)$, $\widetilde{f}(1)$, now correspond to the trace of the Frobenius automorphism on the l-adic cohomology of C (the interesting term $\sum \widetilde{f}(\rho)$ corresponds to the trace on H^1), while the right-hand side corresponds to the number of fixed points of the Frobenius endomorphism, namely the prime divisors of degree 1 on C.

Weil also proved that the Riemann hypothesis is equivalent to the negativity of the right-hand side for all functions $f(x)$ of type

$$f(x) = \int_0^\infty g(xy)\overline{g(y)}dy,$$

whenever $g \in \mathcal{W}$ satisfies the additional conditions

$$\int_0^\infty g(x)\frac{dx}{x} = \int_0^\infty g(x)dx = 0.$$

In the geometric case of curves over a finite field, this negativity is a rather easy consequence of the *algebraic index theorem* for surfaces, namely,

Algebraic Index Theorem. *Let X be a projective nonsingular surface defined over an algebraically closed field. Then the self-intersection quadratic form $(D \cdot D)$, restricted to the group of divisors D on X of degree* 0 *in the projective embedding of X, is negative semidefinite.*

The algebraic index theorem for surfaces is essentially due to Severi[8] in 1906 [**21**, §2,Teo.I]. The proof uses the Riemann–Roch theorem on X and the finiteness of families of curves on X of a given degree; no other proof by algebraic methods is known up to now, although much later several authors independently rediscovered Severi's argument.

The algebraic index theorem for nonsingular projective varieties of even dimension over the complex numbers was first formulated and proved by Hodge, as a consequence of his theory of harmonic forms. No algebraic proof of Hodge's theorem is known, and it remains a fundamental open problem to extend it to the case of varieties over fields of positive characteristic.

The work of Montgomery [**16**], Odlyzko [**17**], and Rudnick and Sarnak [**19**] on correlations for spacings of zeros of $\xi(t)$ suggests that L-functions can be grouped into a few families, in each of which the spacing correlation is universal; the conjectured spacing correlation is the same as for the limiting distribution of eigenvalues of random orthogonal, unitary or symplectic matrices in suitable universal families, as the dimension goes to ∞. All this is compatible with the view expressed by Hilbert and Pólya that the zeros of $\xi(t)$ could be the eigenvalues of a self-adjoint linear operator on an appropriate Hilbert space. It should also be noted that a corresponding

[8]Severi showed that a divisor D on X is algebraically equivalent to 0 up to torsion, if it has degree 0 and $(D \cdot D) = 0$. His proof holds, without modifications, under the weaker assumption $(D \cdot D) \geq 0$, which yields the index theorem.

unconditional theory for the spacing correlations of characteristic roots of Zeta functions of families of algebraic varieties over a finite field has been developed by Katz and Sarnak [**12**], using methods introduced by Deligne in his proof of the Riemann hypothesis for varieties over finite fields. Thus the problem of spacing correlations for zeros of L-functions appears to lie very deep.

All this leads to several basic questions.

Is there a theory in the global case, playing the same role as cohomology does for Zeta functions of varieties over a field of positive characteristic? Is there an analogue of a Frobenius automorphism in the classical case? Is there a general index theorem by which one can prove the classical Riemann hypothesis? We are here in the realm of conjectures and speculation. In the adelic setting propounded by Tate and Weil, the papers [**1**], [**5**], [**7**] offer glimpses of a possible setup for these basic problems.

On the other hand, there are L-functions, such as those attached to Maass waveforms, which do not seem to originate from geometry and for which we still expect a Riemann hypothesis to be valid. For them, we do not have algebraic and geometric models to guide our thinking, and entirely new ideas may be needed to study these intriguing objects.

Bibliography

[1] A. Connes, *Trace formula in noncommutative geometry and the zeros of the Riemann zeta function*, Selecta Math. (NS) **5** (1999), 29–106.

[2] J.B. Conrey, *More than two fifths of the zeros of the Riemann zeta function are on the critical line*, J. reine angew. Math. **399** (1989), 1–26.

[3] P. Deligne, *La conjecture de Weil I*, Publications Math. IHES **43** (1974), 273–308.

[4] P. Deligne, *La conjecture de Weil II*, Publications Math. IHES **52** (1980), 137–252.

[5] C. Deninger, *Some analogies between number theory and dynamical systems on foliated spaces*, in Proc. Int. Congress Math. (Berlin 1998), Vol. I, Doc. Math., Bielefeld, 1998, 163–186.

[6] H.M. Edwards, *Riemann's Zeta Function*, Academic Press, New York, 1974.

[7] S. Haran, *Index theory, potential theory, and the Riemann hypothesis*, in L-functions and Arithmetic, Durham 1990, LMS Lecture Notes **153**, LMS, London, 1991, 257–270.

[8] G.H. Hardy, *Divergent Series*, Oxford Univ. Press, Oxford, 1949, 23–26.

[9] C. Hooley, *On Waring's problem*, Acta Math. **157** (1986), 49–97.

[10] H. Iwaniec and P. Sarnak, *Perspectives on the analytic theory of L-functions*, in Geom. Funct. Analysis, Special Volume, Part II (2000), 705–741.

[11] A. Ivič, *The Riemann Zeta-Function — The Theory of the Riemann Zeta-Function with Applications*, John Wiley, New York, 1985.

[12] N.M. Katz and P. Sarnak, *Random matrices, Frobenius eigenvalues and monodromy*, Amer. Math. Soc. Coll. Publ. **49**, Amer. Math. Soc., Providence, RI, 1999.

[13] E. Landau, *Primzahlen*, Zwei Bd., IInd ed., with an Appendix by Dr. Paul T. Bateman, Chelsea, New York, 1953.

[14] N. Levinson, *More than one-third of the zeros of the Riemann zeta-function are on $\sigma = 1/2$*, Adv. Math. **13** (1974), 383–436.

[15] J. van de Lune, J.J. te Riele, and D.T. Winter, *On the zeros of the Riemann zeta function in the critical strip*, IV, Math. of Comp. **46** (1986), 667–681.

[16] H.L. Montgomery, *Distribution of the zeros of the Riemann zeta function*, in Proceedings Int. Cong. Math. Vancouver 1974, Vol. I, Canad. Math. Congress, Montreal, 1975, 379–381.

[17] A.M. Odlyzko, *Supercomputers and the Riemann zeta function*, in Supercomputing 89: Supercomputing Structures & Computations, Proc. 4-th Intern. Conf. on Supercomputing, International Supercomputing Institute, St. Petersburg, FL, 1989, 348–352.

[18] B. Riemann, *Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse*, in Monat. der Königl. Preuss. Akad. der Wissen. zu Berlin aus der Jahre 1859 (1860), 671–680; also, *Gesammelte math. Werke und wissensch. Nachlass*, 2. Aufl. 1892, 145–155.

[19] Z. Rudnick and P. Sarnak, *Zeros of principal L-functions and random matrix theory*, Duke Math. J. **82** (1996), 269–322.

[20] A. Selberg, *On the zeros of the zeta-function of Riemann*, Der Kong. Norske Vidensk. Selsk. Forhand. **15** (1942), 59–62; also, *Collected Papers*, Vol. I, Springer-Verlag, Berlin, 1989, Vol. I, 156–159.

[21] F. Severi, *Sulla totalità delle curve algebriche tracciate sopra una superficie algebrica*, Math. Annalen **62** (1906), 194–225.

[22] C.L. Siegel, *Über Riemanns Nachlaß zur analytischen Zahlentheorie*, Quellen und Studien zur Geschichte der Mathematik, Astronomie und Physik **2** (1932), 45–80; also *Gesammelte Abhandlungen*, Bd. I, Springer-Verlag, Berlin, 1966, 275–310.

[23] E.C. Titchmarsh, *The Theory of the Riemann Zeta Function*, 2nd ed. revised by R.D. Heath-Brown, Oxford Univ. Press, Oxford, 1986.

[24] R. Taylor and A. Wiles, *Ring theoretic properties of certain Hecke algebras*, Annals Math. **141** (1995), 553–572.

[25] A. Weil, *Œuvres Scientifiques–Collected Papers*, corrected 2nd printing, Vol. I, Springer-Verlag, New York, 1980, 280–298.

[26] A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Hermann & C$^{\text{ie}}$, Paris, 1948.

[27] A. Weil, *Sur les "formules explicites" de la théorie des nombres premiers*, *Meddelanden Från Lunds Univ. Mat. Sem.* (dedié à M. Riesz), (1952), 252–265; also, *Œuvres Scientifiques–Collected Papers*, corrected 2nd printing, Vol. II, Springer-Verlag, New York, 1980, 48–61.

[28] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem*, Annals Math. **141** (1995), 443–551.

Quantum Yang–Mills Theory

Arthur Jaffe and Edward Witten

Quantum Yang–Mills Theory

Arthur Jaffe and Edward Witten

1. The Physics of Gauge Theory

Since the early part of the 20th century, it has been understood that the description of nature at the subatomic scale requires quantum mechanics. In quantum mechanics, the position and velocity of a particle are noncommuting operators acting on a Hilbert space, and classical notions such as "the trajectory of a particle" do not apply.

But quantum mechanics of particles is not the whole story. In 19th and early 20th century physics, many aspects of nature were described in terms of fields—the electric and magnetic fields that enter in Maxwell's equations, and the gravitational field governed by Einstein's equations. Since fields interact with particles, it became clear by the late 1920s that an internally coherent account of nature must incorporate quantum concepts for fields as well as for particles.

After doing this, quantities such as the components of the electric field at different points in space-time become non-commuting operators. When one attempts to construct a Hilbert space on which these operators act, one finds many surprises. The distinction between fields and particles breaks down, since the Hilbert space of a quantum field is constructed in terms of particle-like excitations. Conventional particles, such as electrons, are reinterpreted as states of the quantized field. In the process, one finds the prediction of "antimatter"; for every particle, there must be a corresponding antiparticle, with the same mass and opposite electric charge. Soon after P.A.M. Dirac predicted this on the basis of quantum field theory, the "positron" or oppositely charged antiparticle of the electron was discovered in cosmic rays.

The most important Quantum Field Theories (QFTs) for describing elementary particle physics are gauge theories. The classical example of a gauge theory is Maxwell's theory of electromagnetism. For electromagnetism the gauge symmetry group is the abelian group $U(1)$. If A denotes the $U(1)$ gauge connection, locally a one-form on space-time, then the curvature or

electromagnetic field tensor is the two-form $F = dA$, and Maxwell's equations in the absence of charges and currents read $0 = dF = d * F$. Here $*$ denotes the Hodge duality operator; indeed, Hodge introduced his celebrated theory of harmonic forms as a generalization of the solutions to Maxwell's equations. Maxwell's equations describe large-scale electric and magnetic fields and also—as Maxwell discovered—the propagation of light waves, at a characteristic velocity, the speed of light.

The idea of a gauge theory evolved from the work of Hermann Weyl. One can find in [**34**] an interesting discussion of the history of gauge symmetry and the discovery of Yang–Mills theory [**50**], also known as "non-abelian gauge theory." At the classical level one replaces the gauge group $U(1)$ of electromagnetism by a compact gauge group G. The definition of the curvature arising from the connection must be modified to $F = dA + A \wedge A$, and Maxwell's equations are replaced by the Yang–Mills equations, $0 = d_A F = d_A * F$, where d_A is the gauge-covariant extension of the exterior derivative.

These classical equations can be derived as variational equations from the Yang–Mills Lagrangian

$$L = \frac{1}{4g^2} \int \text{Tr } F \wedge *F, \tag{1}$$

where Tr denotes an invariant quadratic form on the Lie algebra of G. The Yang–Mills equations are nonlinear—in contrast to the Maxwell equations. Like the Einstein equations for the gravitational field, only a few exact solutions of the classical equation are known. But the Yang–Mills equations have certain properties in common with the Maxwell equations: In particular they provide the classical description of massless waves that travel at the speed of light.

By the 1950s, when Yang–Mills theory was discovered, it was already known that the quantum version of Maxwell theory—known as Quantum Electrodynamics or QED—gives an extremely accurate account of electromagnetic fields and forces. In fact, QED improved the accuracy for certain earlier quantum theory predictions by several orders of magnitude, as well as predicting new splittings of energy levels.

So it was natural to inquire whether non-abelian gauge theory described other forces in nature, notably the weak force (responsible among other things for certain forms of radioactivity) and the strong or nuclear force (responsible among other things for the binding of protons and neutrons into nuclei). The massless nature of classical Yang–Mills waves was a serious obstacle to applying Yang–Mills theory to the other forces, for the weak and nuclear forces are short range and many of the particles are massive. Hence these phenomena did not appear to be associated with long-range fields describing massless particles.

In the 1960s and 1970s, physicists overcame these obstacles to the physical interpretation of non-abelian gauge theory. In the case of the weak force, this was accomplished by the Glashow–Salam–Weinberg electroweak theory [**47, 40**] with gauge group $H = SU(2) \times U(1)$. By elaborating the theory with an additional "Higgs field," one avoided the massless nature of classical Yang–Mills waves. The Higgs field transforms in a two-dimensional representation of H; its non-zero and approximately constant value in the vacuum state reduces the structure group from H to a $U(1)$ subgroup (diagonally embedded in $SU(2) \times U(1)$). This theory describes both the electromagnetic and weak forces, in a more or less unified way; because of the reduction of the structure group to $U(1)$, the long-range fields are those of electromagnetism only, in accord with what we see in nature.

The solution to the problem of massless Yang–Mills fields for the strong interactions has a completely different nature. That solution did not come from adding fields to Yang–Mills theory, but by discovering a remarkable property of the quantum Yang–Mills theory itself, that is, of the quantum theory whose classical Lagrangian has been given in (1). This property is called "asymptotic freedom" [**21, 38**]. Roughly this means that at short distances the field displays quantum behavior very similar to its classical behavior; yet at long distances the classical theory is no longer a good guide to the quantum behavior of the field.

Asymptotic freedom, together with other experimental and theoretical discoveries made in the 1960s and 1970s, made it possible to describe the nuclear force by a non-abelian gauge theory in which the gauge group is $G = SU(3)$. The additional fields describe, at the classical level, "quarks," which are spin 1/2 objects somewhat analogous to the electron, but transforming in the fundamental representation of $SU(3)$. The non-abelian gauge theory of the strong force is called Quantum Chromodynamics (QCD).

The use of QCD to describe the strong force was motivated by a whole series of experimental and theoretical discoveries made in the 1960s and 1970s, involving the symmetries and high-energy behavior of the strong interactions. But classical non-abelian gauge theory is very different from the observed world of strong interactions; for QCD to describe the strong force successfully, it must have at the quantum level the following three properties, each of which is dramatically different from the behavior of the classical theory:

(1) It must have a "mass gap;" namely there must be some constant $\Delta > 0$ such that every excitation of the vacuum has energy at least Δ.
(2) It must have "quark confinement," that is, even though the theory is described in terms of elementary fields, such as the quark

fields, that transform non-trivially under $SU(3)$, the physical particle states—such as the proton, neutron, and pion—are $SU(3)$-invariant.

(3) It must have "chiral symmetry breaking," which means that the vacuum is potentially invariant (in the limit, that the quark-bare masses vanish) only under a certain subgroup of the full symmetry group that acts on the quark fields.

The first point is necessary to explain why the nuclear force is strong but short-ranged; the second is needed to explain why we never see individual quarks; and the third is needed to account for the "current algebra" theory of soft pions that was developed in the 1960s.

Both experiment—since QCD has numerous successes in confrontation with experiment—and computer simulations, see for example [**8**], carried out since the late 1970s, have given strong encouragement that QCD does have the properties cited above. These properties can be seen, to some extent, in theoretical calculations carried out in a variety of highly oversimplified models (like strongly coupled lattice gauge theory, see, for example, [**48**]). But they are not fully understood theoretically; there does not exist a convincing, whether or not mathematically complete, theoretical computation demonstrating any of the three properties in QCD, as opposed to a severely simplified truncation of it.

2. Quest for Mathematical Understanding

In surveying the physics of gauge theories in the last section, we considered both classical properties—such as the Higgs mechanism for the electroweak theory—and quantum properties that do not have classical analogs—like the mass gap and confinement for QCD. Classical properties of gauge theory are within the reach of established mathematical methods, and indeed classical non-abelian gauge theory has played a very important role in mathematics in the last twenty years, especially in the study of three- and four-dimensional manifolds. On the other hand, one does not yet have a mathematically complete example of a quantum gauge theory in four-dimensional space-time, nor even a precise definition of quantum gauge theory in four dimensions. Will this change in the 21st century? We hope so!

At times, mathematical structures of importance have first appeared in physics before their mathematical importance was fully recognized. This happened with the discovery of calculus, which was needed to develop Newtonian mechanics, with functional analysis and group representation theory, topics whose importance became clearer with quantum mechanics, and even with the study of Riemannian geometry, whose development was greatly intensified once it became clear, through Einstein's invention of General

Relativity to describe gravity, that this subject plays a role in the description of nature. These areas of mathematics became generally accessible only after a considerable time, over which the ideas were digested, simplified, and integrated into the general mathematical culture.

Quantum Field Theory (QFT) became increasingly central in physics throughout the 20th century. There are reasons to believe that it may be important in 21st century mathematics. Indeed, many mathematical subjects that have been actively studied in the last few decades appear to have natural formulations—at least at a heuristic level—in terms of QFT. New structures spanning probability, analysis, algebra, and geometry have emerged, for which a general mathematical framework is still in its infancy.

On the analytic side, a byproduct of the existence proofs and mathematical construction of certain quantum field theories was the construction of new sorts of measures, in particular non-Gaussian, Euclidean-invariant measures on spaces of generalized functionals. Dirac fields and gauge fields require measures on spaces of functions taking values in a Grassmann algebra and on spaces of functions into other target geometries.

Renormalization theory arises from the physics of quantum field theory and provides a basis for the mathematical investigation of local singularities (ultra-violet regularity) and of global decay (infra-red regularity) in quantum field theories. Asymptotic freedom ensures a decisive regularity in the case when classical Sobolev inequalities are borderline. Surprisingly, the ideas from renormalization theory also apply in other areas of mathematics, including classic work on the convergence of Fourier series and recent progress on classical dynamical systems.

On the algebraic side, investigations of soluble models of quantum field theory and statistical mechanics have led to many new discoveries involving topics such as Yang–Baxter equations, quantum groups, Bose–Fermi equivalence in two dimensions, and rational conformal field theory.

Geometry abounds with new mathematical structures rooted in quantum field theory, many of them actively studied in the last twenty years. Examples include Donaldson theory of 4-manifolds, the Jones polynomial of knots and its generalizations, mirror symmetry of complex manifolds, elliptic cohomology, and $SL(2,\mathbb{Z})$ symmetry in the theory of affine Kac–Moody algebras.

QFT has in certain cases suggested new perspectives on mathematical problems, while in other cases its mathematical value up to the present time is motivational. In the case of the geometric examples cited above, a mathematical definition of the relevant QFTs (or one in which the relevant physical techniques can be justified) is not yet at hand. Existence theorems that put QFTs on a solid mathematical footing are needed to make the

geometrical applications of QFT into a full-fledged part of mathematics. Regardless of the future role of QFT in pure mathematics, it is a great challenge for mathematicians to understand the physical principles that have been so important and productive throughout the twentieth century.

Finally, QFT is the jumping-off point for a quest that may prove central in 21st century physics—the effort to unify gravity and quantum mechanics, perhaps in string theory. For mathematicians to participate in this quest, or even to understand the possible results, QFT must be developed further as a branch of mathematics. It is important not only to understand the solution of specific problems arising from physics, but also to set such results within a new mathematical framework. One hopes that this framework will provide a unified development of several fields of mathematics and physics, and that it will also provide an arena for the development of new mathematics and physics.

For these reasons the Scientific Advisory Board of CMI has chosen a Millennium problem about quantum gauge theories. Solution of the problem requires both understanding one of the deep unsolved physics mysteries, the existence of a mass gap, and also producing a mathematically complete example of quantum gauge field theory in four-dimensional space-time.

3. Quantum Fields

A quantum field, or local quantum field operator, is an operator-valued generalized function on space-time obeying certain axioms. The properties required of the quantum fields are described at a physical level of precision in many textbooks, see, for example, [**27**]. Gårding and Wightman gave mathematically precise axioms for quantum field theories on $\mathbb{R}^4$ with a Minkowski signature, see [**45**], and Haag and Kastler introduced a related scheme for local functions of the field, see [**24**].

Basically, one requires that the Hilbert space $\mathcal{H}$ of the quantum field carry a representation of the Poincaré group (or inhomogeneous Lorentz group). The Hamiltonian H and momentum $\vec{P}$ are the self-adjoint elements of the Lie algebra of the group that generate translations in time and space. A *vacuum vector* is an element of $\mathcal{H}$ that is invariant under the (representation of the) Poincaré group. One assumes that the representation has positive energy, $0 \leq H$, and a vacuum vector $\Omega \in \mathcal{H}$ that is unique up to a phase. Gauge-invariant functions of the quantum fields also act as linear transformations on $\mathcal{H}$ and transform covariantly under the Poincaré group. Quantum fields in space-time regions that cannot be connected by a light signal should be independent; Gårding and Wightman formulate independence as the commuting of the field operators (anti-commuting for two fermionic fields).

One of the achievements of 20th century axiomatic quantum field theory was the discovery of how to convert a Euclidean-invariant field theory on a Euclidean space-time to a Lorentz-invariant field theory on Minkowski space-time, and vice-versa. Wightman used positive energy to establish analytic continuation of expectations of Minkowski field theories to Euclidean space. Kurt Symanzik interpreted the Euclidean expectations as a statistical mechanical ensemble of classical Markov fields [**46**], with a probability density proportional to $\exp(-S)$, where S denotes the Euclidean action functional. E. Nelson reformulated Symanzik's picture and showed that one can also construct a Hilbert space and a quantum-mechanical field from a Markov field [**33**]. Osterwalder and Schrader then discovered the elementary "reflection-positivity" condition to replace the Markov property. This gave rise to a general theory establishing equivalence between Lorentzian and Euclidean axiom schemes [**35**]. See also [**13**].

One hopes that the continued mathematical exploration of quantum field theory will lead to refinements of the axiom sets that have been in use up to now, perhaps to incorporate properties considered important by physicists such as the existence of an operator product expansion or of a local stress-energy tensor.

4. The Problem

To establish existence of four-dimensional quantum gauge theory with gauge group G, one should define a quantum field theory (in the above sense) with local quantum field operators in correspondence with the gauge-invariant local polynomials in the curvature F and its covariant derivatives, such as $\operatorname{Tr} F_{ij}F_{kl}(x)$.[1] Correlation functions of the quantum field operators should agree at short distances with the predictions of asymptotic freedom and perturbative renormalization theory, as described in textbooks. Those predictions include among other things the existence of a stress tensor and an operator product expansion, having prescribed local singularities predicted by asymptotic freedom.

Since the vacuum vector Ω is Poincaré invariant, it is an eigenstate with zero energy, namely $H\Omega = 0$. The positive energy axiom asserts that in any quantum field theory, the spectrum of H is supported in the region $[0, \infty)$. A quantum field theory has a *mass gap* Δ if H has no spectrum in the interval $(0, \Delta)$ for some $\Delta > 0$. The supremum of such Δ is the mass m, and we require $m < \infty$.

[1] A natural 1–1 correspondence between such classical 'differential polynomials' and quantized operators does not exist, since the correspondence has some standard subtleties involving renormalization [**27**]. One expects that the space of classical differential polynomials of dimension $\leq d$ does correspond to the space of local quantum operators of dimension $\leq d$.

YANG–MILLS EXISTENCE AND MASS GAP. *Prove that for any compact simple gauge group G, a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ and has a mass gap $\Delta > 0$. Existence includes establishing axiomatic properties at least as strong as those cited in* [**45, 35**].

5. Comments

An important consequence of the existence of a mass gap is clustering: Let $\vec{x} \in \mathbb{R}^3$ denote a point in space. We let H and $\vec{P}$ denote the energy and momentum, generators of time and space translation. For any positive constant $C < \Delta$ and for any local quantum field operator $\mathcal{O}(\vec{x}) = e^{-i\vec{P}\cdot\vec{x}}\mathcal{O}e^{i\vec{P}\cdot\vec{x}}$ such that $\langle\Omega, \mathcal{O}\Omega\rangle = 0$, one has

$$|\langle\Omega, \mathcal{O}(\vec{x})\mathcal{O}(\vec{y})\Omega\rangle| \leq \exp(-C|\vec{x} - \vec{y}|), \tag{2}$$

as long as $|\vec{x} - \vec{y}|$ is sufficiently large. Clustering is a locality property that, roughly speaking, may make it possible to apply mathematical results established on $\mathbb{R}^4$ to any 4-manifold, as argued at a heuristic level (for a supersymmetric extension of four-dimensional gauge theory) in [**49**]. Thus the mass gap not only has a physical significance (as explained in the introduction), but it may also be important in mathematical applications of four-dimensional quantum gauge theories to geometry. In addition the existence of a uniform gap for finite-volume approximations may play a fundamental role in the proof of existence of the infinite-volume limit.

There are many natural extensions of the Millennium problem. Among other things, one would like to prove the existence of an isolated one-particle state (an upper gap, in addition to the mass gap), to prove confinement, to prove existence of other four-dimensional gauge theories (incorporating additional fields that preserve asymptotic freedom), to understand dynamical questions (such as the possible mass gap, confinement, and chiral symmetry breaking) in these more general theories, and to extend the existence theorems from $\mathbb{R}^4$ to an arbitrary 4-manifold.

But a solution of the existence and mass gap problem as stated above would be a turning point in the mathematical understanding of quantum field theory, with a chance of opening new horizons for its applications.

6. Mathematical Perspective

Wightman and others have questioned for approximately fifty years whether mathematically well-defined examples of relativistic, nonlinear quantum field theories exist. We now have a partial answer: Extensive results on the existence and physical properties of nonlinear QFTs have been proved through the emergence of the body of work known as "constructive quantum field theory" (CQFT).

The answers are partial, for in most of these field theories one replaces the Minkowski space-time $\mathbb{M}^4$ by a lower-dimensional space-time $\mathbb{M}^2$ or $\mathbb{M}^3$, or by a compact approximation such as a torus. (Equivalently in the Euclidean formulation one replaces Euclidean space-time $\mathbb{R}^4$ by $\mathbb{R}^2$ or $\mathbb{R}^3$.) Some results are known for Yang–Mills theory on a 4-torus $\mathbb{T}^4$ approximating $\mathbb{R}^4$, and, while the construction is not complete, there is ample indication that known methods could be extended to construct Yang–Mills theory on $\mathbb{T}^4$.

In fact, at present we do not know any non-trivial relativistic field theory that satisfies the Wightman (or any other reasonable) axioms in four dimensions. So even having a detailed mathematical construction of Yang–Mills theory on a compact space would represent a major breakthrough. Yet, even if this were accomplished, no present ideas point the direction to establish the existence of a mass gap that is uniform in the volume. Nor do present methods suggest how to obtain the existence of the infinite volume limit $\mathbb{T}^4 \to \mathbb{R}^4$.

6.1. Methods. Since the inception of quantum field theory, two central methods have emerged to show the existence of quantum fields on non-compact configuration space (such as Minkowski space). These known methods are

(i) Find an exact solution in closed form;
(ii) Solve a sequence of approximate problems, and establish convergence of these solutions to the desired limit.

Exact solutions may be available for nonlinear fields for special values of the coupling which yields extra symmetries or integrable models. They might be achieved after clever changes of variables. In the case of four-dimensional Yang–Mills theory, an exact solution appears unlikely, though there might some day be an asymptotic solution in a large N limit.

The second method is to use mathematical approximations to show the convergence of approximate solutions to exact solutions of the nonlinear problems, known as *constructive quantum field theory*, or CQFT. Two principle approaches—studying quantum theory on Hilbert space, and studying classical functional integrals—are related by the Osterwalder–Schrader construction. Establishing uniform *a priori* estimates is central to CQFT, and three schemes for establishing estimates are

(a) correlation inequalities,
(b) symmetries of the interaction,
(c) convergent expansions.

The correlation inequality methods have an advantage; they are general. But correlation inequalities rely on special properties of the interaction that

often apply only for scalar bosons or abelian gauge theories. The use of symmetry also applies only to special values of the couplings and is generally combined with another method to obtain analytic control. In most known examples, perturbation series, i.e., power series in the coupling constant, are divergent expansions; even Borel and other resummation methods have limited applicability.

This led to development of expansion methods, generally known as *cluster expansions*. Each term in a cluster expansion sum depends on the coupling constants in a complicated fashion; they often arise as functional integrals. One requires sufficient quantitative knowledge of the properties of each term in an expansion to ensure the convergence of the sum and to establish its qualitative properties. Refined estimates yield the rate of exponential decay of Green's functions, magnitude of masses, the existence of symmetry breaking (or its preservation), etc.

Over the past thirty years, a panoply of expansion methods have emerged as a central tool for establishing mathematial results in CQFT. In their various incarnations, these expansions encapsulate ideas of the asymptotic nature of perturbation theory, of space-time localization, of phase-space localization, of renormalization theory, of semi-classical approximations (including "non-perturbative" effects), and of symmetry breaking. One can find an introduction to many of these methods and references in [**18**], and a more recent review of results in [**28**]. These expansion methods can be complicated and the literature is huge, so we can only hope to introduce the reader to a few ideas; we apologize in advance for important omissions.

6.2. The First Examples: Scalar Fields. Since the 1940s the quantum Klein–Gordon field φ provided an example of a linear, scalar, mass-m field theory (arising from a quadratic potential). This field, and the related free spinor Dirac field, served as models for formulating the first axiom schemes in the 1950s [**45**].

Moments of a Euclidean-invariant, reflection-positive, ergodic, Borel measure $d\mu$ on the space $\mathcal{S}'\left(\mathbb{R}^d\right)$ of tempered distributions may satisfy the Osterwalder–Schrader axioms. The Gaussian measure $d\mu$ with mean zero and covariance $C = (-\Delta + m_0^2)^{-1}$ yields the free, mass-m_0 field; but one requires non-Gaussian $d\mu$ to obtain nonlinear fields. (For the Gaussian measure, reflection positivity is equivalent to positivity of the transformation ΘC, restricted to $L^2(\mathbb{R}^d_+) \subset L^2(\mathbb{R}^d)$. Here $\Theta : t \to -t$ denotes the time-reflection operator, and $\mathbb{R}^d_+ = \{(t, \vec{x}) : t \geq 0\}$ is the positive-time subspace.)

The first proof that nonlinear fields satisfy the Wightman axioms and the first construction of such non-Gaussian measures only emerged in the 1970s. The initial examples comprised fields with small, polynomial nonlinearities on $\mathbb{R}^2$: first in finite volume, and then in the infinite volume limit [**15, 19,**

22]. These field theories obey the Wightman axioms (as well as all other axiomatic formulations), the fields describe particles of a definite mass, and the fields produce multi-particle states with non-trivial scattering [**19**]. The scattering matrix can be expanded as an asymptotic series in the coupling constants, and the results agree term-by-term with the standard description of scattering in perturbation theory that one finds in physics texts [**37**].

A quartic Wightman QFT on $\mathbb{R}^3$ also exists, obtained by constructing a remarkable non-Gaussian measure $d\mu$ on $\mathcal{S}'(\mathbb{R}^3)$ [**16, 10**]. This merits further study.

We now focus on some properties of the simplest perturbation to the action-density of the free field, namely, the even quartic polynomial

$$\lambda\varphi^4 + \frac{1}{2}\sigma\varphi^2 + c. \tag{3}$$

The coefficients $0 < \lambda$ and $\sigma, c \in \mathbb{R}$ are real parameters, all zero for the free field. For $0 < \lambda \ll 1$, one can choose $\sigma(\lambda), c(\lambda)$ so the field theory described by (3) exists, is unique, and has a mass equal to m such that $|m - m_0|$ is small.

Because of the local singularity of the nonlinear field, one must first cut off the interaction. The simplest method is to truncate the Fourier expansion of the field φ in (3) to $\varphi_\kappa(x) = \int_{|k|\leq\kappa} \widetilde{\varphi}(k)e^{-ikx}dk$ and to compactify the spatial volume of the perturbation to $\mathcal{V}$. One obtains the desired quantum field theory as a limit of the truncated approximations. The constants σ, c have the form $\sigma = \alpha\lambda + \beta\lambda^2$ and $c = \gamma\lambda + \delta\lambda^2 + \epsilon\lambda^3$. One desires that the expectations of products of fields have a limit as $\kappa \to \infty$. One chooses α, γ (in dimension 2), and one chooses all the coefficients $\alpha, \beta, \gamma, \delta, \epsilon$ (in dimension 3), to depend on κ in the way that perturbation theory suggests. One then proves that the expectations converge as $\kappa \to \infty$, even though the specified constants $\alpha, \ldots$ diverge. These constants provide the required renormalization of the interaction. In the three-dimensional case one also needs to normalize vectors in the Fock space a constant that diverges with κ; one calls this constant a wave-function renormalization constant.

The "mass" operator in natural units is $M = \sqrt{H^2 - \vec{P}^2} \geq 0$, and the vacuum vector Ω is a null vector, $M\Omega = 0$. Massive single particle states are eigenvectors of an eigenvalue $m > 0$. If m is an isolated eigenvalue of M, then one infers from the Wightman axioms and Haag–Ruelle scattering theory that asymptotic scattering states of an arbitrary number of particles exist, see [**24, 18**].

The fundamental problem of *asymptotic completeness* is the question whether these asymptotic states (including possible bound states) span $\mathcal{H}$. Over the past thirty years, several new methods have emerged, yielding proofs of asymptotic completeness in scattering theory for non-relativistic

quantum mechanics. This gives some hope that one can now attack the open problem of asymptotic completeness for any known example of nonlinear quantum field theory.

In contrast to the existence of quantum fields with a φ^4 nonlinearity in dimensions 2 and 3, the question of extending these results to four dimensions is problematic. It is known that self-interacting scalar fields with a quartic nonlinearity do not exist in dimension 5 or greater [**12, 1**]. (The proofs apply to field theories with a single, scalar field.) Analysis of the borderline dimension 4 (between existence and non-existence) is more subtle; if one makes some reasonable (but not entirely proved) assumptions, one also can conclude triviality for the quartic coupling in four dimensions. Not only is this persuasive evidence, but furthermore the quartic coupling does not have the property of asymptotic freedom in four dimensions. Thus all insights from random walks, perturbation theory, and renormalization analysis point toward triviality of the quartic interaction in four dimensions.

Other polynomial interactions in four dimensions are even more troublesome: The classical energy of the cubic interaction is unbounded from below, so it appears an unlikely candidate for a quantum theory where positivity of the energy is an axiom. And polynomial interactions of degree greater than quartic are more singular in perturbation theory than the quartic interaction.

All these reasons complement the physical and geometric importance of Yang–Mills theory and highlight it as the natural candidate for four-dimensional CQFT.

6.3. Large Coupling Constant. In two dimensions, the field theory with energy density (3) exists for all positive λ. For $0 \leq \lambda \ll 1$ the solution is unique under a variety of conditions; but for $\lambda \gg 1$ two different solutions exist, each characterized by its ground state or "phase." The solution in each phase satisfies the Osterwalder–Schrader and Wightman axioms with a non-zero mass gap and a unique, Poincaré-invariant vacuum state. The distinct solutions appear as a bifurcation of a unique approximating solution with finite volume $\mathcal{V}$ as $\mathcal{V} \to \infty$.

One exhibits this behavior by reordering and scaling the $\lambda\varphi^4$ interaction (3) with $\lambda \gg 1$ to obtain an equivalent double-well potential of the form

$$\lambda \left(\varphi^2 - \frac{1}{\lambda} \right)^2 + \frac{1}{2}\sigma\varphi^2 + c. \tag{4}$$

Here $\lambda \ll 1$ is a new coupling constant and the renormalization constants σ and c are somewhat different from those above. The two solutions for a given λ are related by the broken $\varphi \to -\varphi$ symmetry of the interaction (4). The proof of these facts relies upon developing a convergent cluster

expansion about each minimum of the potential arising from (4) and proving the probability of tunneling between the two solutions is small [**20**].

A critical value λ_c of λ in (3) provides a boundary between the uniqueness of the solution (for $\lambda < \lambda_c$) and the existence of a phase transition $\lambda > \lambda_c$. As λ increases to λ_c, the mass gap $m = m(\lambda)$ decreases monotonically and continuously to zero [**23, 17, 32**].

The detailed behavior of the field theory (or the mass) in the neighborhood of $\lambda = \lambda_c$ is extraordinarily difficult to analyze; practically nothing has been proved. Physicists have a qualitative picture based on the assumed fractional power-law behavior $m(\lambda) \sim |\lambda_c - \lambda|^\nu$ above or below the critical point, where the exponent ν depends on the dimension. One also expects that the critical coupling λ_c corresponds to the greatest physical force between particles, and that these critical theories are close to scaling limits of Ising-type modes in statistical physics. One expects that further understanding of these ideas will result in new computational tools for quantum fields and for statistical physics.

There is some partial understanding of a more general multi-phase case. One can find an arbitrary number n of phases by making a good choice of a polynomial energy density $\mathcal{P}_n(\varphi)$ with n minima. It is interesting to study the perturbation of a fixed such polynomial $\mathcal{P}_n$ by polynomials $\mathcal{Q}$ of lower degree and with small coefficients. Among these perturbations one can find families of polynomials $\mathcal{Q}(\varphi)$ that yield field theories with exactly $n' \le n$ phases [**26**].

6.4. Yukawa Interactions and Abelian Gauge Theory. The existence of boson-fermion interactions is also known in two dimensions, and partial results exist in three dimensions. In two dimensions Yukawa interactions of the form $\overline{\psi}\psi\varphi$ exist with appropriate renormalization, as well as their generalizations of the form $\mathcal{P}(\varphi) + \overline{\psi}\psi\mathcal{Q}''(\varphi)$, see [**42, 18**]. The supersymmetric case $\mathcal{P} = |\mathcal{Q}'|^2$ requires extra care in dealing with cancellations of divergences, see [**28**] for references.

A continuum two-dimensional Higgs model describes an abelian gauge field interacting with a charged scalar field. Brydges, Fröhlich, and Seiler constructed this theory and showed that it satisfies the Osterwalder–Schrader axioms [**7**], providing the only complete example of an interacting gauge theory satisfying the axioms. A mass gap exists in this model [**4**]. Extending all these conclusions to a non-abelian Higgs model, even in two dimensions, would represent a qualitative advance.

Partial results on the three-dimensional $\overline{\psi}\psi\varphi$ interaction have been established, see [**30**], as well as for other more singular interactions [**14**].

6.5. Yang–Mills Theory. Much of the mathematical progress reviewed above results from understanding functional integration and using those methods to construct Euclidean field theories. Functional integration for gauge theories raises new technical problems revolving around the rich group of symmetries, especially gauge symmetry. Both the choice of gauge and the transformation between different choices complicate the mathematical structure; yet gauge symmetry provides the possibility of asymptotic freedom. Certain insights and proposals in the physics literature [**9, 5**] have led to an extensive framework; yet the implications of these ideas for a mathematical construction of Yang–Mills theory need further understanding.

Wilson suggested a different approach based on approximating continuum space-time by a lattice, on which he defined a gauge-invariant action [**48**]. With a compact gauge group and a compactified space-time, the lattice approximation reduces the functional integration to a finite-dimensional integral. One must then verify the existence of limits of appropriate expectations of gauge-invariant observables as the lattice spacing tends to zero and as the volume tends to infinity.

Reflection positivity holds for the Wilson approximation [**36**], a major advantage; few methods exist to recover reflection positivity in case it is lost through regularization—such as with dimensional regularization, Pauli–Villiars regularization, and many other methods. Establishing a quantum mechanical Hilbert space is part of the solution to this Millennium problem.

Balaban studied this program in a three-dimensional lattice with periodic boundary conditions, approximating a space-time torus [**2**]. He studied renormalization transformations (integration of large-momentum degrees of freedom followed by rescaling) and established many interesting properties of the effective action they produce. These estimates are uniform in the lattice spacing, as the spacing tends to zero. The choices of gauges are central to this work, as well as the use of Sobolev space norms to capture an analysis of geometric effects.

One defines these gauges in phase cells: The choices vary locally in space-time, as well as on different length scales. The choices evolve inductively as the renormalization transformations proceed, from gauges suitable for local regularity (ultraviolet gauges) to those suitable for macroscopic distances (infrared gauges). This is an important step toward establishing the existence of the continuum limit on a compactified space-time. These results need to be extended to the study of expectations of gauge-invariant functions of the fields.

While this work in three dimensions is important in its own right, a qualitative breakthrough came with Balaban's extension of this analysis to four dimensions [**3**]. This includes an analysis of asymptotic freedom to

control the renormalization group flow as well as obtaining quantitative estimates on effects arising from large values of the gauge field.

Extensive work has also been done on a continuum regularization of the Yang–Mills interaction, and it has the potential for further understanding [**39, 29**].

These steps toward understanding quantum Yang–Mills theory lead to a vision of extending the present methods to establish a complete construction of the Yang–Mills quantum field theory on a compact, four-dimensional space-time. One presumably needs to revisit known results at a deep level, simplify the methods, and extend them.

New ideas are needed to prove the existence of a mass gap that is uniform in the volume of space-time. Such a result presumably would enable the study of the limit as $\mathbb{T}^4 \to \mathbb{R}^4$.[2]

6.6. Further Remarks. Because four-dimensional gauge theory is a theory in which the mass gap is not classically visible, to demonstrate it may require a non-classical change of variables or "duality transformation." For example, duality has been used to establish a mass gap in the statistical mechanics problem of a Coulomb gas, where the phenomenon is known as Debye screening: Macroscopic test charges in a neutral Coulomb gas experience a mutual force that decays exponentially with the distance. The mathematical proof of this screening phenomenon proceeds through the identity of the partition function of the Coulomb gas to that of a $\cos(\lambda\varphi)$ (sine-Gordon) field theory, and the approximate parabolic potential near a minimum of this potential, see [**6**].

One view of the mass gap in Yang–Mills theory suggests that it could arise from the quartic potential $(A \wedge A)^2$ in the action, where $F = dA + gA \wedge A$, see [**11**], and may be tied to curvature in the space of connections, see [**44**]. Although the Yang–Mills action has flat directions, certain quantum mechanics problems with potentials involving flat directions (directions for which the potential remains bounded as $|x| \to \infty$) do lead to bound states [**43**].

A prominent speculation about a duality that might shed light on dynamical properties of four-dimensional gauge theory involves the $1/N$ expansion [**25**]. It is suspected that four-dimensional quantum gauge theory with gauge group $SU(N)$ (or $SO(N)$, or $Sp(N)$) may be equivalent to a string theory with $1/N$ as the string coupling constant. Such a description might give a clear-cut explanation of the mass gap and confinement and perhaps a good starting point for a rigorous proof (for sufficiently large N).

[2]We specifically exclude weak-existence (compactness) as the solution to the existence part of the Millennium problem, unless one also uses other techniques to establish properties of the limit (such as the existence of a mass gap and the axioms).

There has been surprising progress along these lines for certain strongly coupled four-dimensional gauge systems with matter [**31**], but as of yet there is no effective approach to the gauge theory without fermions. Investigations of supersymmetric theories and string theories have uncovered a variety of other approaches to understanding the mass gap in certain four-dimensional gauge theories with matter fields; for example, see [**41**].

Bibliography

[1] M. Aizenman, *Geometric analysis of φ^4 fields and Ising models*, Comm. Math. Phys. **86** (1982), 1–48.

[2] T. Balaban, *Ultraviolet stability of three-dimensional lattice pure gauge field theories*, Comm. Math. Phys. **102** (1985), 255–275.

[3] T. Balaban, *Renormalization group approach to lattice gauge field theories. I: generation of effective actions in a small field approximation and a coupling constant renormalization in 4D*, Comm. Math. Phys. **109** (1987), 249–301.

[4] T. Balaban, D. Brydges, J. Imbrie, and A. Jaffe, *The mass gap for Higgs models on a unit lattice*, Ann. Physics **158** (1984), 281–319.

[5] C. Becchi, A. Rouet, and R. Stora, *Renormalization of gauge theories*, Ann. Phys. **98** (1976), 287–321.

[6] D. Brydges and P. Federbush, *Debye screening*, Comm. Math. Phys. **73** (1980), 197–246.

[7] D. Brydges, J. Fröhlich, and E. Seiler, *On the construction of quantized gauge fields, I.* Ann. Phys. **121** (1979), 227–284, II. Comm. Math. Phys. **71** (1980), 159–205, and III. Comm. Math. Phys. **79** (1981), 353–399.

[8] M. Creutz, *Monte carlo study of quantized $SU(2)$ gauge theory*, Phys. Rev. **D21** (1980), 2308–2315.

[9] L. D. Faddeev and V. N. Popov, *Feynman diagrams for the Yang–Mills fields*, Phys. Lett. **B25** (1967), 29–30.

[10] J. Feldman and K. Osterwalder, *The Wightman axioms and the mass gap for weakly coupled φ^4_3 quantum field theories*, Ann. Physics **97** (1976), 80–135.

[11] R. Feynman, *The quantitative behavior of Yang–Mills theory in $2+1$ dimensions*, Nucl. Phys. **B188** (1981), 479–512.

[12] J. Fröhlich, *On the triviality of $\lambda\varphi^4_d$ theories and the approach to the critical point*, Nucl. Phys. **200** (1982), 281–296.

[13] J. Fröhlich, K. Osterwalder, and E. Seiler, *On virtual representations of symmetric spaces and theory analytic continuation*, Ann. Math. **118** (1983), 461–489.

[14] K. Gawedzki, *Renormalization of a non-renormalizable quantum field theory*, Nucl. Phys. **B262** (1985), 33–48.

[15] J. Glimm and A. Jaffe, *The $\lambda\varphi^4_2$ quantum field theory without cut-offs, I.* Phys. Rev. **176** (1968), 1945–1951, *II.* Ann. Math. **91** (1970), 362–401, *III.* Acta. Math. **125** (1970), 203–267, and *IV.* J. Math. Phys. **13** (1972), 1568–1584.

[16] J. Glimm and A. Jaffe, *Positivity of the φ^4_3 Hamiltonian*, Fortschr. Phys. **21** (1973), 327–376.

[17] J. Glimm and A. Jaffe, *φ^4_2 quantum field model in the single phase region: differentiability of the mass and bounds on critical exponents*, Phys. Rev. **10** (1974), 536–539.

[18] J. Glimm and A. Jaffe, *Quantum Physics*, Second Edition, Springer Verlag, 1987, and *Selected Papers, Volumes I and II*, Birkhäuser Boston, 1985. (Volume II includes reprints of [**15, 16, 19, 20**].)

[19] J. Glimm, A. Jaffe, and T. Spencer, *The Wightman axioms and particle structure in the $P(\varphi)_2$ quantum field model*, Ann. Math. **100** (1974), 585–632.

[20] J. Glimm, A. Jaffe, and T. Spencer, *A convergent expansion about mean field theory, I.* Ann. Phys. **101** (1976), 610–630, and *II.* Ann. Phys. **101** (1976), 631–669.
[21] D. J. Gross and F. Wilczek, *Ultraviolet behavior of non-abelian gauge theories*, Phys. Rev. Lett. **30** (1973), 1343–1346.
[22] F. Guerra, L. Rosen, and B. Simon, *The $P(\varphi)_2$ Euclidean quantum field theory as classical statistical mechanics*, Ann. Math. **101** (1975), 111–259.
[23] F. Guerra, L. Rosen, and B. Simon, *Correlation inequalities and the mass gap in $P(\varphi)_2$, III. Mass gap for a class of strongly coupled theories with non-zero external field*, Comm. Math. Phys. **41** (1975), 19–32.
[24] R. Haag, *Local Quantum Physics*, Springer Verlag, 1992.
[25] G. 't Hooft, *A planar diagram theory for strong interactions*, Nucl. Phys. **B72** (1974), 461–473.
[26] J. Imbrie, *Phase diagrams and cluster expansions for low temperature $\mathcal{P}(\varphi)_2$ models*, Comm. Math. Phys. **82** (1981), 261–304 and 305–343.
[27] C. Itzykson and J.-B. Zuber, *Quantum Field Theory*, McGraw-Hill, New York, 1980.
[28] A. Jaffe, *Constructive quantum field theory*, in Mathematical Physics, T. Kibble, ed., World Scientific, Singapore, 2000.
[29] J. Magnen, Vincent Rivasseau, and Roland Sénéor, *Construction of YM_4 with an infrared cutoff*, Comm. Math. Phys. **155** (1993), 325–383.
[30] J. Magnen and R. Sénéor, *Yukawa quantum field theory in three dimensions*, in Third International Conference on Collective Phenomena, J. Lebowitz et. al. eds., New York Academy of Sciences, 1980.
[31] J. Maldacena, *The large N limit of superconformal field theories and supergravity*, Adv. Theor. Math. Phys. **2** (1998), 231–252.
[32] O. McBryan and J. Rosen, *Existence of the critical point in φ^4 field theory*, Comm. Math. Phys. **51** (1976), 97–105.
[33] E. Nelson, *Quantum fields and Markoff fields*, in Proc. Sympos. Pure Math. XXIII 1971, AMS, Providence, R.I., 1973, 413–420.
[34] L. O'Raifeartaigh, *The Dawning of Gauge Theory*, Princeton University Press, 1997.
[35] K. Osterwalder and R. Schrader, *Axioms for Euclidean Green's functions*, Comm. Math. Phys. **31** (1973), 83–112, and Comm. Math. Phys. **42** (1975), 281–305.
[36] K. Osterwalder and E. Seiler, *Gauge theories on the lattice*, Ann. Phys. **110** (1978), 440–471.
[37] K. Osterwalder and R. Sénéor, *A nontrivial scattering matrix for weakly coupled $P(\varphi)_2$ models*, Helv. Phys. Acta **49** (1976), 525–535.
[38] H. D. Politzer, *Reliable perturbative results for strong interactions?* Phys. Rev. Lett. **30** (1973), 1346–1349.
[39] V. Rivasseau, *From Perturbative to Constructive Renormalization*, Princeton Univ. Press, Princeton, NJ, 1991.
[40] A. Salam, Weak and electromagnetic interactions, in Svartholm: Elementary Particle Theory, Proceedings of The Nobel Symposium held in 1968 at Lerum, Sweden, Stockholm, 1968, 366–377.
[41] N. Seiberg and E. Witten, *Electric-magnetic duality, monopole condensation, and confinement in $N = 2$ supersymmetric Yang–Mills theory*, Nucl. Phys. **B426** (1994), 19–52.
[42] E. Seiler, *Schwinger functions for the Yukawa model in two dimensions*, Comm. Math. Phys. **42** (1975), 163–182.
[43] B. Simon, *Some quantum operators with discrete spectrum but classically continuous spectrum*, Ann. Phys. **146** (1983), 209–220.
[44] I. M. Singer, *The geometry of the orbit space for nonabelian gauge theories*, Phys. Scripta **24** (1981), 817–820.
[45] R. Streater and A. Wightman, *PCT, Spin and Statistics and all That*, W. A. Benjamin, New York, 1964.

[46] K. Symanzik, *Euclidean quantum field theory*, in Local Quantum Theory, R. Jost, ed., Academic Press, New York, 1969, 152–226.
[47] S. Weinberg, *A model of leptons*, Phys. Rev. Lett. **19** (1967), 1264–1266.
[48] K. G. Wilson, *Quarks and strings on a lattice*, in New Phenomena In Subnuclear Physics, Proceedings of the 1975 Erice School, A. Zichichi, ed., Plenum Press, New York, 1977.
[49] E. Witten, *Supersymmetric Yang–Mills theory on a four-manifold*, J. Math. Phys. **35** (1994), 5101–5135.
[50] C. N. Yang and R. L. Mills, *Conservation of isotopic spin and isotopic gauge invariance*, Phys. Rev. **96** (1954), 191–195.

Rules for the Millennium Prizes

The Clay Mathematics Institute (CMI) of Cambridge, Massachusetts, has named seven "Millennium Prize Problems". The Scientific Advisory Board of CMI (SAB) selected these problems, focusing on important classic questions that have resisted solution over the years. The Board of Directors of CMI designated a US$7 million prize fund for the solution to these problems, with US$1 million allocated to each. The directors of CMI, and no other persons or body, have the authority to authorize payment from this fund or to modify or interpret these stipulations. The Board of Directors of CMI makes all mathematical decisions for CMI, upon the recommendation of its SAB.

The SAB of CMI will consider a proposed solution to a Millennium Prize Problem if it is a complete mathematical solution to one of the problems. (In the case that someone discovers a mathematical counterexample, rather than a proof, the question will be considered separately as described below.) A proposed solution to one of the Millennium Prize Problems may not be submitted directly to CMI for consideration.

Before consideration, a proposed solution must be published in a refereed mathematics publication of worldwide repute (or such other form as the SAB shall determine qualifies), and it must also have general acceptance in the mathematics community two years after. Following this two-year waiting period, the SAB will decide whether a solution merits detailed consideration. In the affirmative case, the SAB will constitute a special advisory committee, which will include (a) at least one SAB member and (b) at least two non-SAB members who are experts in the area of the problem. The SAB will seek advice to determine potential non-SAB members who are internationally recognized mathematical experts in the area of the problem. As part of this procedure, each component of a proposed solution under consideration shall be verified by one or more members of this special advisory committee.

The special advisory committee will report within a reasonable time to the SAB. Based on this report and possible further investigation, the SAB will make a recommendation to the Directors. The SAB may recommend the award of a prize to one person. The SAB may recommend that a particular prize be divided among multiple solvers of a problem or their heirs. The SAB

will pay special attention to the question of whether a prize solution depends crucially on insights published prior to the solution under consideration. The SAB may (but need not) recommend recognition of such prior work in the prize citation, and it may (but need not) recommend the inclusion of the author of prior work in the award.

If the SAB cannot come to a clear decision about the correctness of a solution to a problem, its attribution, or the appropriateness of an award, the SAB may recommend that no prize be awarded for a particular problem. If new information comes to light, the SAB may (but will not necessarily) reconsider a negative decision to recommend a prize for a proposed solution, but only after an additional two-year waiting period following the time that the new information comes to light. The SAB has the sole authority to make recommendations to the directors of the CMI concerning the appropriateness of any award and the validity of any claim to the CMI Millennium Prize.

In the case of the **P** versus **NP** problem and the Navier–Stokes problem, the SAB will consider the award of the Millennium Prize for deciding the question in either direction. In the case of the other problems, if a counterexample is proposed, the SAB will consider the counterexample after publication, and the same two-year waiting period as for a proposed solution will apply. If, in the opinion of the SAB, the counterexample effectively resolves the problem, then the SAB may recommend the award of the Prize. If the counterexample shows that the original problem survives after reformulation or elimination of some special case, then the SAB may recommend that a small prize be awarded to the author. The money for this prize will not be taken from the Millennium Prize Problem fund, but from other CMI funds.

Any person who is not a disqualified person (as that term is defined in section 4946 of the Internal Revenue Code) in connection with the institute may receive the Millennium Prize. Any disqualified person other than a substantial contributor to the institute (as defined in section 507 of the Internal Revenue Code) may also receive the Millennium Prize provided that the directors, upon application for the prize by a disqualified person, shall modify the procedures outlined herein for selecting an awardee so as to assure that the candidate is not present during and does not participate in any deliberations of the Board, the SAB, or any special award committee in connection with making the award and provided further that if an award is made to a disqualified person, the Board shall make public the procedures that are adopted to assure impartiality and to avoid conflict of interest. For purposes of this paragraph, members of the SAB shall be considered "disqualified persons".

With the one exception in the prior paragraph, all decision-making procedures concerning the CMI Millennium Prize Problems are private. This

includes the deliberations or recommendations of any person or persons CMI has used to obtain advice on this question. No records of these deliberations or related correspondence may be made public without the prior approval of the directors, the SAB, and all other living persons involved, unless fifty years of time have elapsed after the event in question.

Please send inquiries regarding the Millennium Prize Problems to prize.problems@claymath.org.

Authors' Biographies

Enrico Bombieri was born in Italy in 1940. He studied number theory with Giovanni Ricci in Milan and Harold Davenport in Cambridge, England, and algebraic geometry with Aldo Andreotti in Pisa. He currently is Professor of Mathematics at the Institute for Advanced Study in Princeton, New Jersey.

His research spans number theory, complex analysis, partial differential equations, algebraic geometry, and algebra. In 1974 he was awarded the Fields Medal for his work in mathematics and in particular for his work on Linnik's large sieve and the solution of Bernstein's problem on minimal surfaces.

Stephen Cook was born in Buffalo, New York, received his B.Sc. degree from the University of Michigan in 1961 and his S.M. and Ph.D. degrees from Harvard University in 1962 and 1966 respectively. From 1966 to 1970 he was Assistant Professor at the University of California, Berkeley. He joined the faculty at the University of Toronto in 1970 as an Associate Professor and was promoted to Professor in 1975 and to University Professor in 1985.

Cook's principal research areas are computational complexity and proof complexity, with excursions into programming language semantics and parallel computation. He is the author of over 60 research papers, including his 1971 paper "The Complexity of Theorem Proving Procedures," which introduced the theory of NP completeness and proved that the Boolean satisfiability problem is NP complete. He is the 1982 recipient of the Turing Award. He is a Fellow of the Royal Society of London, Royal Society of Canada, and was elected to membership in the National Academy of Sciences (United States) and the American Academy of Arts and Sciences. Twenty-six students have completed their Ph.D. degrees under his supervision.

Pierre Deligne learned algebraic geometry from A. Grothendieck. He was a permanent member of the Institut des Hautes Études Scientifiques from 1970 to 1984, and has been a Professor at the Institute for Advanced Study since 1985. Deligne was awarded the Fields Medal in 1978 for his solution of the Weil conjectures.

One of Deligne's recurrent interests has been the panoply of cohomologies attached to algebraic varieties: the analogies between them, the structures they carry, and their applications. Algebraic cycles, or substitutes thereof, play a central role in our present understanding of the relations between cohomology theories.

Charles Fefferman received his B.S. from the University of Maryland and his Ph.D. from Princeton (1969) under E.M. Stein. After four years at the University of Chicago, Fefferman returned to Princeton, where he has been ever since. He has worked on classical Fourier analysis, partial differential equations, several complex variables, quantum mechanics, conformal geometry, fluid mechanics and computational geometry. He has served as an editor of the *Annals of Mathematics* and the *Proceedings of the National Academy of Sciences*. He also served as chairman of the Princeton mathematics department. His honors include the Alan Waterman award and the Fields Medal. He is a member of the National Academy of Sciences, the American Academy of Arts and Sciences, and the American Philosophical Society.

Jeremy Gray was born in 1947 and studied mathematics at Oxford and Warwick University, where he took his doctorate in 1980. He has taught at the Open University since 1974, where he is now a professor of the History of Mathematics and Director of the Centre for the History of the Mathematical Sciences. Gray is also an Honorary Professor in the University of Warwick Mathematics Department.

Gray is the author, co-author, or editor of 13 books and is presently working on two more. He works on the history of mathematics in the 19th and 20th centuries, and also on issues in the philosophy and cultural significance of mathematics. His most recent published book is *János Bolyai, non-Euclidean geometry, and the Nature of Space*, a Burndy Library publication, MIT Press, 2004. He has recently finished a book on the history of geometry in the 19th century and is now working on one on mathematical modernism and the philosophy of mathematics.

Arthur Michael Jaffe is a mathematician and physicist known for using insight from physics to widen the frontiers of mathematics, as well as for developing new mathematical tools to prove results relevant for theoretical physics. His work solved an old problem by showing the mathematical compatibility of quantum theory, special relativity, and particle interaction. He laid the foundations and established many of the mathematical results proving existence and establishing properties of non-linear quantum field theories in two-dimensional and three-dimensional space-time — much in collaboration with James Glimm. The proofs include constructing the only known non-Gaussian, Euclidean-invariant, reflection-positive measures on

the space of generalized functions on two-dimensional and three-dimensional Euclidean space. Solving the "mass gap" problem for these examples played a central role. This work became the centerpiece of an active school of mathematical physics known as "constructive quantum field theory." A positive solution to the millennium challenge question on "Yang–Mills" theory would represent the natural extension of constructive field theory to four-dimensional space-time.

Jaffe has also worked in other areas of mathematics, including classical gauge theories, non-commutative geometry, and relations between analysis, geometry, and quantum field theory. He served as President of the American Mathematical Society, and he shaped the foundation of the Clay Mathematics Institute.

John Milnor was born in Orange, New Jersey, in 1931. He spent his undergraduate and graduate student years at Princeton University, working on knot theory under the supervision of Ralph Fox and dabbling in game theory with his fellow student John Nash. However, like his mathematical grandfather, Solomon Lefschetz, he had great difficulty sticking to one subject. Under the inspiration of Norman Steenrod and John Moore, he branched out into algebraic and differential toplogy. This led to problems in pure algebra, including algebraic K-theory, and the study of quadratic forms. More recently, conversations with William Thurston and Adrien Douady led to problems in real and complex dynamical systems, which have occupied him for twenty years or so. Still restless, one current activity is an attempt to study problems of complexity in the life sciences.

After many years in Princeton, at the University and also at the Institute for Advanced Study, and after shorter stays at UCLA and MIT, Milnor moved to Stony Brook in 1980 as director of a small Institute for Mathematical Sciences. Milnor was awarded the Fields Medal in 1962 for his work in topology, including his discovery that the 7-sphere can have several differential structures.

Andrew Wiles is one of the world's preeminent number theorists, celebrated for his proof of Fermat's Last Theorem and noted for his articles on Iwasawa theory, elliptic curves, modular forms, and Galois representations. Wiles holds the Eugene Higgins Professorship of Mathematics at Princeton University. He has received numerous awards and honors including the Shaw Prize in 2005 from the Shaw Foundation of Hong Kong and the Wolf Prize in 1996 from the Israeli Wolf Foundation. He is a member of the National Academy of Sciences (USA) and of the Académie des Sciences, Paris. Elected a Fellow of the Royal Society in 1989, Wiles was knighted by the Queen of England in 2000.

Edward Witten received his B.A. from Brandeis University in 1971 and his M.A. and Ph.D. in Physics from Princeton University in 1974 and 1976, respectively. Witten was a postdoctoral fellow from 1976 to 1977, and then a Junior Fellow at Harvard University from 1977 to 1980. In September 1980, Witten was appointed professor of Physics at Princeton. He was awarded a MacArthur Fellowship in 1982. In 1987 he was appointed Professor in the School of Natural Sciences at the Institute for Advanced Study. Edward Witten's work has had profound impact, both on fundamental theories of physics and on mathematics. In 1990, his contributions to mathematics were recognized at the International Congress of Mathematics, when Witten was awarded the Fields Medal.