Mathematics at the turn of the millennium*

Phillip A. Griffiths

This century has been a golden age for mathematics, especially for so-called 'pure' mathematics, where the questions we study are internal to the field. Many important, long-standing problems have been resolved, both by gaining a richer and deeper understanding of the structure of the subfields of mathematics, and by exploring the interactions between subfields. Now, at the turn of the century, the interactivity between subfields is expanding toward interactivity between mathematics and the other sciences. These interactions have led both to great insights within the sciences, and to the broadening and deepening of mathematics itself.

The world of mathematics

In discussing our subject, we mathematicians face a dilemma. The most effective way to explain mathematics to general readers is to use metaphors, which entails a loss of precision and carries the risk of misunderstanding. On the other hand, advanced mathematical terms are incomprehensible to most people – including other scientists. As my colleague David Mumford, president of the International Mathematics Union, has said, 'I am accustomed, as a professional mathematician, to living in a sort of vacuum, surrounded by people who ... declare with an odd sort of pride that they are mathematically illiterate'.

Within the mathematical community, however, the use of a precise language is a distinct advantage. Because of its abstract nature and universality, mathematics knows neither linguistic nor political boundaries. It is one reason that mathematics has always carried a distinctly international flavour. A mathematician in Japan can usually read the paper of a colleague in Germany without translation.

The number of highly active research mathematical scientists worldwide is small – probably well under 10,000 – so that a given subfield may be populated by a small number of highly specialized individuals. By necessity, these colleagues know one another other well, regardless of country of residence, and collaborate over long distances. During the present century, a growing number of papers have been co-authored by mathematicians from different nations (the number rose by about 50% between 1981 and 1993). And so mathematicians are well adapted to the current trend toward a world of vanishing borders.

But what is it that these mathematicians do? In general, mathematics can be described as the search for structures and patterns that bring order and simplicity to our universe. It may be said that the object or beginning point of a mathematical study is not as important as the patterns and coherence that emerge. These patterns and coherence often add to the power of mathematics by bringing clarity to a completely different object or process – to another branch of mathematics, another science, or to society at large.

When mathematicians speak of their work, two words carry great importance. Mathematics is a field where a 'problem' is not a bad thing. In fact, a good problem is what mathematicians yearn for; it signifies interesting work. The second word is 'proof', which strongly suggests the rigour of the discipline. Arthur Eddington once said, 'Proof is an idol before which the mathematician tortures himself'. A mathematical proof is a formal and logical line of reasoning that begins with a set of axioms and moves through logical steps to a conclusion. A proof, once given, is permanent; some have existed since the time of the Greeks. A proof confirms truth for the mathematician the way experiment or observation does for the natural scientist.

The 20th century has been a fertile time for the resolution of long-standing problems, and for a wealth of accomplishments that would require at least an encyclopedia to describe. Let us look at just two of the more interesting achievements – proofs to problems that were over 300 years old. Both occurred towards the end of the century and could have succeeded only because of the mathematics that preceded them.

Fermat's Last Theorem

The first is the solution of Fermat's Last Theorem by Andrew Wiles, which made news around the globe in 1993. This example is interesting because of Fermat, an eccentric jurist and amateur mathematician who pub-

*Presented at the Conference 'Frontiers of the Mind in the 21st Century', Library of Congress, Washington DC, 15 June 1999; Also presented at Jawaharlal Nehru Centre for Advanced Scientific Research and Indian Institute of Science, Bangalore as Newton lecture.

Phillip A. Griffiths is at the Institute of Advanced Study, Olden Lane, Princeton, NJ 08540, USA.

lished no papers; because of Wiles, who toiled on the problem alone for seven years; and because of the problem itself, whose solution depended on fundamental advances in number theory by many mathematicians over a period of 350 years, especially during the last half-century.

The theorem was written in 1637, when Pierre de Fermat was studying an ancient Greek text on number theory called *Arithmetica*, by Diofantus. Interest in number theory had waned since the time of the ancient Greeks, but Fermat loved numbers. He came across the famous Pythagorean equation most of us learn in school: $x^2 + y^2 = z^2$. Even today, countless school children learn to say, 'The square of the hypotenuse equals the sum of the squares of the other two sides'.

Of particular interest are solutions to the Pythagorean equation in whole numbers, such as the beautiful 3–4–5 right triangle. When Fermat saw this, he noted that for any exponent greater than two, the equation could not have solutions in whole numbers. He also wrote, in Latin, that he had discovered his own wonderful proof, but that the margin was too small to contain it. No such proof has ever been found. Fermat made many such marginal questions (some regarded as taunts to his fellow mathematicians), and over the centuries they were all answered except this one, Fermat's Last Theorem.

Andrew Wiles first came across Fermat at age 10, in a library in Cambridge, England, where he grew up. He vowed that some day he would prove it. By the time he was a young mathematician, however, he had learned that pursuing Fermat by itself was not advisable, and decided to work instead in a complex area of algebraic number theory known as Iwasawa theory. But he never forgot about Fermat.

In 1986, he learned of a breakthrough: a colleague named Ken Ribet at the University of California at Berkeley had linked Fermat's Last Theorem to another unsolved problem, the Taniyama–Shimura conjecture, a surprising and brilliant formulation in algebraic geometry posed in 1955. To summarize a very complex sequence of reasoning, this linkage showed that proving the Taniyama–Shimura conjecture would essentially prove Fermat's Last Theorem as well. It constructed a logical bridge between the intricate worlds of elliptical curves and modular forms, a kind of dictionary that allows questions and insights to be translated between the two worlds. It also meant that Wiles' earlier work in algebraic number theory would be helpful, and that he would probably generate some interesting problems – whether or not he found a proof.

He did find a proof – after a series of baffling obstacles and sudden insights. Even after he had presented his results, a small but crucial error was found during the refereeing process, which lead to a further year's work. Again, there seemed to be no solution – and again, there was one. Wiles called this last insight 'the most important moment of my working life. It was so indescribably beautiful, it was so simple and elegant and I just stared in disbelief for twenty minutes'.

Did Fermat really complete his own proof in the 17th century? Undoubtedly, some will continue to look for evidence that he did, but it is highly unlikely. Wiles' work made use of whole subfields of 19th and 20th century mathematics that did not exist in Fermat's time. Beneath Fermat's equation now lies an enormous and elaborate formal structure – the kind of structure that mathematicians strive for. The solution to Fermat arises from the implications of understanding that structure.

'Perhaps I can best describe my experience of doing mathematics in terms of a journey through a dark unexplored mansion. You enter the first room of the mansion and it is completely dark. You stumble around bumping into the furniture, but gradually you learn where each piece of furniture is. Finally, after six months or so, you find the light switch, you turn it on, and suddenly it is all illuminated. You can see exactly where you were. Then you move into the next room and spend another six months in the dark. So each of these breakthroughs, while sometimes they are momentary, sometimes over a period of a day or two, they are the culmination of – and could not exist without – the many months of stumbling around in the dark that proceed them.'

Andrew Wiles, who proved Fermat's Last Theorem in 1993

Kepler's sphere packing conjecture

The second problem is Kepler's sphere packing conjecture. Like the Fermat problem, sphere packing could only have been solved in the way that it was in the last few decades. Even so, it took Thomas Hales, professor of mathematics at the University of Michigan, ten years to do so. Like Fermat, sphere packing sounds simple, but it defeated mathematicians for nearly four centuries. Moreover, both problems had subtle difficulties that led countless mathematicians to believe they had found solutions that turned out to be false.

The question was posed in the latter half of the 16th century, when Walter Raleigh asked the English mathematician Thomas Harriot for a quick way to estimate the number of cannon balls that could be stacked on the deck of a ship. In turn, Harriot wrote to Johannes Kepler, the German astronomer, who was already interested in stacking: how could spheres be arranged to minimize the gaps among them? Kepler could find no

system more efficient than the way sailors naturally stack cannon balls, or grocers stack oranges: face-centered cubic packing. Kepler declared that by this technique 'the packing will be the tightest possible, so that in no other arrangement could more pellets be stuffed into the same container'. This assertion became known as the Kepler conjecture, but it was not proved.

Major progress was made in the 19th century, when the legendary German mathematician Karl Friedrich Gauss proved that the orange-pile arrangement was the most efficient among all 'lattice packings', but this did not rule out the possibility of a more efficient non-lattice arrangement. By the end of the 19th century the Kepler conjecture was deemed sufficiently important for another famous mathematician, David Hilbert, to include in his list of 23 great 'turn-of-the-century' problems.

The problem is difficult because of the immense number of possibilities that must be eliminated. By the mid-20th century, mathematicians had discovered how to reduce it to a finite problem, but this problem was too complex to compute. A major advance came in 1953 when the Hungarian mathematician Laszlo Fejes Tóth reduced the problem to an immense calculation involving many specific cases and suggested how it might be solved by computer.

Even for Hales, with modern computers, the challenge was immense. His equation has 150 variables, each of which must be changed to describe every conceivable stacking arrangement. The proof, explained in a 250-page argument which contains 3 gigabytes of computer files, relies extensively on methods from the theory of global optimization, linear programming, and interval arithmetic. Hales acknowledged that for a proof so long and complex, it would be some time before anyone could confirm all its details.

It is worth noting, however, that his exercise was far from frivolous. The topic of sphere packing belongs to a crucial part of the mathematics that lies behind the error-detecting and error-correcting codes that are widely used to store information on compact disks and to compress information for transmission around the world. In today's information society, it is difficult to think of a more significant application. And let's not forget the grocers around the world, who can now trust that they *do* stack oranges in the most efficient way possible.

The four-colour problem

As an addendum to sphere-packing, a related problem is worth mentioning – the so-called four-colour problem of map-making. This is the assertion that only four colours are needed to colour any map so that no neighbouring countries have the same colour. This problem is similar to sphere packing: it is an elementary problem which probably seemed straightforward when first proposed by the English mathematician Francis Guthrie in 1852. It is also similar in that the existing proof reduces it to a finite problem which then required heavy amounts of computing capability.

The proof, accomplished in 1976 by Wolfgang Haken and Kenneth Appel, involved showing that if you can test a list of x maps, then the test is true for all maps. Although the number of conceivable maps is infinite, Haken and Appel showed that the colourability of all of them depends only on the colourability of a large but finite set of fundamental maps. This was the first significant problem to succumb to the raw power of the computer. At the same time, it has caused some people to suggest that 'brute-force' computer proofs lack the clarity of traditional proof: that is, they prove that the conjecture is true, but do not explain why. We may expect a good deal of further debate on this point.

The dual nature of mathematics

The three proofs I have just mentioned could be described as intellectual exercises of great precision, abstraction, and, some would say, beauty. Indeed, the mathematician G. H. Hardy once said that the practice of mathematics can only be justified as an art form. In fact, there *is* a parallel with the arts here: Mathematicians, like artists, place great value on the aesthetic quality of their work.

But I want to suggest that mathematics has a dual nature, and that this is part of the reason for its vitality. In addition to its intellectual and aesthetic qualities, mathematics is tremendously useful in the real world. Earlier this century, the physicist Eugene Wigner spoke famously of 'the unreasonable effectiveness of mathematics'. To mention just a few classical examples of this effectiveness, the modern computer was made possible by Liebnitz' binary number code; Einstein formulated his Theory of Relativity with the help of Riemannian geometry; and the edifices of quantum mechanics, crystallography, and communications technology all rest firmly on the platform of group theory. Nowadays, one would add effectiveness in drug design, economics, finance, telecommunications, and many other fields.

Mathematicians have always carried their discoveries into adjacent fields where they have produced new insights and whole new subfields. Francis Bacon, at the dawn of the Enlightenment in 1605, prefigured this principle of integrative science with an apt image: 'No perfect discovery can be made upon a flat or a level: neither is it possible to discover the more remote or deeper parts of any science, if you stand but upon the level of the same science and ascend not to a higher science'.

Repeatedly in the 20th century, mathematics has ascended to that higher place. For example, the develop-

ment of X-ray tomography (the CAT and MRI scanning technologies) was built upon integral geometry; the generation of codes for secure transmission of data depends on the arithmetic of prime numbers; and the design of large, efficient networks in telecommunications uses infinite-dimensional representations of groups.

Thus, mathematics is both an independent discipline valued for precision and intrinsic beauty, and it is a rich source of tools for application in the 'real' world. And the two parts of this duality are intimately connected. As we shall see in the following section, it is the strengthening of these connections during the 20th century that had allowed the field to steadily gain effectiveness – both within mathematics and with the world beyond.

Trends of the 20th century

A principal reason mathematics is healthy today is the breakdown of barriers within the field. At first glance, the full span of mathematics – an enormous body of concepts, conjectures, hypotheses, and theorems amassed over more than 2000 years – seems to defy the possibility of unity. Gone are the days when a single giant – an Euler or a Gauss – could command its entirety. With the rapid development of subfields after World War II, mathematics became so specialized that practitioners had difficulty communicating with anyone outside their own speciality. And today these specialists are commonly scattered between Bonn, Princeton, Berkeley, and Tokyo.

But this trend toward fragmentation is complemented by the growing tendency to address interesting problems in an overarching manner. Subfields, once viewed as quite disjoint, are now seen as part of a whole, as new connections emerge between them. For example, algebraic geometry, the field I am most familiar with, is a field that combines algebra, geometry, topology and analysis. As we near the end of this century, synergies in this strongly interconnected area have played a major role in some of the crowning achievements of pure mathematics. One, of course, is the solution of Fermat's Last Theorem, which was mentioned earlier. Another is the solution of the Mordell Conjecture, which states that a polynomial equation with rational coefficients of degree 4 or more can have at most a finite number of rational solutions (the Fermat equation has no such solutions). A third is solution of the Weil conjectures, which are the analogs for finite fields of the Riemann hypothesis (discussed later). All these accomplishments reflect the ability of mathematicians simultaneously to draw on multiple subfields and to perceive their subject as a whole.

Solitons

One of the most remarkable achievements of mathematics of the latter half of the 20th century is the theory of solitons, which illustrates the underlying unity of the field. These can be described as nonlinear waves that exhibit extremely unexpected and interesting behaviour.

First, a bit of background. Traditionally, we talk about two different kinds of waves. The first are linear waves, which are familiar in everyday life, such as light waves or sound waves. Linear waves have several characteristics: One, they move uniformly through space without changing. That is, they have a constant velocity, no matter what their shape; C-sharp travels at the same speed as E-flat. And they have a constant amplitude; a C-sharp remains C-sharp if you hear it a block away. The second characteristic is the property of superposition: if you play multiple notes on the piano simultaneously, you always hear the sum of all the notes at once, which brings us harmony. Even a very complicated sound can be resolved into its constituent harmonics.

The second kind of waves is nonlinear waves, which are less familiar and quite different. A simple example can be seen as an ocean wave approaches the shore. The amplitude, wavelength, and velocity, which are constant in linear waves, all can be seen to change. The distance between the wave tops decreases, the height increases as the waves 'feel' the bottom, and the velocity changes; the upper part overtakes the bottom part and falls over it as the wave breaks. In an even more intricate event, two waves may come together, interact in a complicated, nonlinear way, and give rise to three outgoing waves instead of two.

Now we come to solitons. The story begins in 1834, when a Scottish engineer named John Scott Russell was trying to determine the most efficient design for canal boats. One day, he observed that waves in a shallow canal sometimes behave in very peculiar fashion. Some waves would travel at a constant velocity without changing shape, but those with large amplitude went faster than those with small amplitude. A large wave might overtake a smaller one, resulting in a complex interaction, whereupon the large wave would emerge travelling faster than the smaller one. After this nonlinear interaction, they would again act like linear waves.

In the middle of the 20th century, a group of mathematicians were studying a nonlinear wave equation. Because it described nonlinear waves, they expected that its solution would develop singularities, or breaks, at some point. They wrote a computer program to numerically solve the equation and found that the wave did not break as expected. This led them to look at the Korteweg–de Vries equation, which was written down a century ago to describe the behaviour of waves in shallow water. It was found that the phenomena observed by

Lord Russell were provable mathematically for the Korteweg–de Vries equation; in other words, the solution to that equation exhibited soliton behaviour. These are extremely unusual equations, because solitons are in some ways like linear waves and in other ways like nonlinear waves.

This discovery provoked a rush of activity which exhibited in the most beautiful way the unity of mathematics. It involved developments in computation, and in mathematical analysis, which is the traditional way to study differential equations. It also turns out that one can understand the solutions to these differential equations through certain very elegant constructions in algebraic geometry. Additionally, the solutions are intimately related to representation theory, in that these equations turn out to have an infinite number of hidden symmetries.

Finally, they relate back to problems in elementary geometry. For example, an interesting problem is to find the surface of a cone with a fixed volume but least area among all surfaces with a given boundary. It is not at all evident that this has anything to do with a shallow water wave, but in fact it does. The differential equations that describe the solution turn out to have soliton behaviour in the same way as the equations describing shallow water waves. So we have started with two mathematical problems – one in mathematical physics and one in differential geometry – and found that in each of them, the same rich, underlying structure leads to extremely rare and interesting soliton behaviour.

Mathematics and the other sciences

Beyond the breakdown of internal barriers, mathematics has become much more interactive with other sciences and with business, finance, security, management, decision-making, and the modelling of complex systems. And some of these disciplines, in turn, are challenging mathematicians with interesting new types of problems which then lead to new applications.

Mathematics and theoretical physics. Mathematics has been linked with theoretical physics for centuries, and the linkage has grown stronger over the last two decades. For example, algebraic geometry has become an essential tool for theoretical physicists in their search for a unified field theory – more precisely, for a theory that unifies gravity with the three fundamental forces of physics: the strong nuclear force, the weak nuclear force and electromagnetism.

One interesting candidate for a new unifying theory is string theory, which is pursued at my own institution. The name comes from the proposal that the most elementary building blocks of matter are tiny, vibrating loops or segments that are string-like in shape and vibrate in many different modes, like violin strings. The effort to understand this extremely complex theory has led a group of theoretical physicists deep into mathematics, where they have made a series of spectacular predictions about mathematics; these predictions are beginning to be verified. The results have stimulated a flurry of work that continues to add to the plausibility of the theory; it has also spawned a new branch of four-dimensional mathematics called quantum geometry which, in turn, is opening new insights to physics.

Another indication of the close relationship between mathematics and physics was seen in the 1998 awards of Fields Medals, the highest honour in mathematics. Of the four medallists, three of them worked in areas with a strong physics influence, and a special award was given for work in quantum computing, whose roots are in quantum mechanics.

Mathematics and the life sciences. One of the fastest-growing new partnerships is the collaboration between mathematics and biology. The partnership began in the field of ecology in the 1920s, when the Italian mathematician Vito Volterra developed the first models of predator–prey relationships. He found that the waxing and waning of predator and prey populations of fish could best be described mathematically. After World War II, the modelling methods developed for populations were extended to epidemiology, which resembles population biology in being the study of diseases in large populations of people.

Most recently, the insights of molecular genetics have inspired scientists to adapt these same methods to infectious diseases, where the objects of study are not populations of organisms or people, but populations of cells. In a cellular system, the predator is a population of viruses, for example, and the prey is a population of human cells. These two populations rise and fall in a complex Darwinian struggle for survival that lends itself to mathematical description. In the last decade, the ability to use mathematical models that describe infectious agents as predators and host cells as prey has redefined many aspects of immunology, genetics, epidemiology, neurology and drug design. The reason this partnership is successful is that mathematical models offer the first tools of sufficient power to describe the immensity of numbers and relationships found in biological systems.

For example, mathematical biologists have been able to make quantitative predictions about how viruses and other microbes grow in their hosts, how they change the genetic structure of hosts, and how they interact with the host's immune system. Some of the most surprising results have emerged in the study of the AIDS epidemic, reversing our understanding of HIV viruses in infected patients. The prevailing view had been that HIV viruses

lie dormant for a period of 10 or so years before beginning to infect host cells and cause disease. Mathematical modelling has shown that the HIV viruses that cause the most diseases are not dormant; they grow steadily and rapidly, with a half-life of only about 2 days.

Why, then does it take an average of 10 years for infection to begin? Again, mathematical modelling has shown that disease progression may be caused by viral evolution. The immune system is capable of suppressing the virus for a long time, but eventually new forms of viruses mutate and become abundant and overwhelm the immune defence. This happens because viruses, like other infectious agents, can reproduce faster than their hosts, and the reproduction of their genetic material is less accurate. Virtually every HIV infection is seen as an evolutionary process in which the virus population constantly changes and new virus mutants continuously emerge. Natural selection favours variants able to escape the immune response, or to infect more kinds of cells in the human body, or to reproduce faster. The models show that all evolutionary changes are those that increase virus abundance in the patient and thereby accelerate disease.

These same mathematical models have brought an understanding of why anti-HIV drugs should be given in combination, and given as early as possible during infection. They are most effective in combination because viruses seldom produce multiple mutations at once. And they should be given early before viral evolution can progress very far.

A major threat to human health in the next century will be microbial resistance to drug therapy, an area to which mathematical models can contribute. Models can point to clear guidelines for collection and analysis of data, which can make drugs more effective. Good models of the complex interactions between infectious agents and the immune system can eventually create a new discipline of quantitative immunology.

There are many more new partnerships between mathematics and the other sciences; much of the most innovative and productive work is being done at the frontiers between fields and disciplines. An excellent example is the study of fluid dynamics, where mathematics interacts with fields such as meteorology, anatomy and geology. Describing the complex movement of fluids – hurricanes, blood flow thorough the heart, oil in porous ground – was virtually impossible before the discovery that a purely mathematical construct called the Navier–Stokes equations can do just that, and today these equations are applied to fluid theory in various disciplines. Another example is control theory, a branch of the theory of dynamical systems. In just one application, much of the testing of high-performance aircraft can now be done by computer simulations, greatly reducing the expense and danger of wind tunnels and test flights.

It is important to emphasize that while modelling and simulation are modern and important topics, we are still not very good at addressing the uncertainties that are present in these complex simulations. Learning to grapple with uncertainty is high on the list of priorities for mathematicians, who must develop fundamentally new approaches if they are to understand how uncertainties arise in models and how they propagate through systems. Our models will only be as accurate as our ability to smooth out their uncertainties.

Research challenges for the 21st century

Despite the tremendous achievements of the 20th century, dozens of outstanding mathematical problems await solution. Most mathematicians probably agree that the following three examples are among the most challenging and interesting.

The Riemann Hypothesis

The first is the Riemann Hypothesis, which has tantalized mathematicians for 150 years. It has to do with the concept of prime numbers, which are the basic building blocks of arithmetic. A prime number, as many people remember from high school, is a positive whole number greater than 1 that cannot be divided by any other positive number except by itself and 1. The primes begin with 2, 3, 5, 7, 11 and 13, and rise from there toward infinity. As long ago as the third century BC, Euclid suggested that no one could ever find the 'largest' prime number; in other words, they are infinite in number. And the known primes are growing ever larger: in 1992, a French team found a prime with 227,832 digits.

For someone studying prime numbers by pencil and paper, they appear at first to occur randomly. But in the 19th century, the German mathematician Bernhard Riemann extended Euclid's observation to assert that primes are not only infinite in number, but that they should occur in a very subtle and precise pattern. Proving – or disproving – this is perhaps the deepest existing problem in pure mathematics.

The Poincaré Conjecture

We return to Henri Poincaré, who suggested a problem that is baffling both because it is so fundamental and because it seems so simple. In Poincaré's days, a century ago, it was even regarded as trivial, as was the whole field of topology – a field which he essentially invented. But today topology is a vital and significant subfield of mathematics.

In rough terms, topology is concerned with the fundamental properties of structures and spaces. A sphere,

for example, can be stretched, compressed or warped in any number of ways (in a topologist's eyes) and still remain a sphere, as long as it is not torn or punctured. A topologist sees a donut and a coffee cup as identical, because either one of them can be massaged into the same basic shape as the other – a ring with a hole in it, or torus. Of special interest to topologists are manifolds, which mean 'having multiple features or forms'. A soccer ball, for example, is a two-dimensional manifold, or two-sphere; we can manipulate it any way we want as long as we do not rip it, and it will still be a soccer ball.

Topologists seek to identify all possible manifolds, including the shape of the universe – which is the subject of the Poincaré Conjecture. This is relatively easy in two dimensions, and was done by the end of the 19th century. The test of whether a manifold is really a two-sphere is also straightforward. Imagine placing a rubber band on the surface of a soccer ball. If the rubber band can be shrunk to a point without leaving the surface, and if this can be done anywhere on the surface, the ball is a two-sphere, and we say it is simply connected.

In 1904, Poincaré conjectured that what is true for two dimensions is also true for three – that any simply-connected, three-dimensional manifold (such as the universe) has to be a three-sphere. This sounds intuitively obvious, but nobody has ever shown that there are not some false three-spheres, so the conjecture still has not been proven. Surprisingly, proofs are known for the equivalent of Poincaré conjecture for all dimensions strictly greater than 3, but not yet for 3.

Does P = NP?

A third major problem is related to the philosophical question of what is knowable and what is unknowable. In 1931, the Austrian-born logician Kurt Gödel established that complete certainty could not be found in arithmetic – assuming that arithmetic is founded on certain 'self-evident' properties, or axioms, of whole numbers. And in the theory of computing, Alan Turing set down rules in the 1930s to decide what is computable and what is not. A more refined question is to ask what is computable in polynomial time, or P time.

A familiar example of P/NP concerns the travelling salesman who needs to visit n number of cities. When n is a small number, one can write a computer program to compute the best route in a reasonable amount of computer time, or polynomial time. As n becomes larger, computer time may increase exponentially, until the problem becomes computationally intractable, or NP. Most numerical codes today are designed on the assumption that factoring them is NP, a computationally intractable problem. This assumption has enormous implications for safe use of the Internet, where large numbers are used as security codes.

In fact there are some very interesting current developments on the 'P vs NP' question which may be related to the Gödel incompleteness theorem mentioned above. It seems possible that certain mathematical statements that eventually include lower bounds on computation, such as 'P does not equal NP', cannot be proved within the framework of Peano arithmetic, or set theory. (Peano arithmetic is the standard, or most natural, version of arithmetic.)

This thesis is not yet proved, but its resolution appears feasible in the foreseeable future. What is known is that, first, all techniques used so far to prove lower bounds on computational models reside in a specific low fragment of Peano arithmetic; and second, proved techniques in these fragments cannot separate P from NP – unless the integer factorization has much faster algorithms than we currently know or suspect. In other words, whether a problem is P or NP will depend on whether or not we can factor integers much faster than we have thought possible.

Theoretical computer science

The field we are discussing, theoretical computer science, is one of the most important and active areas of scientific study today. It was actually founded half a century ago, before computers existed, when Alan Turning and his contemporaries set out to mathematically define the concept of 'computation' and to study its power and limits. These questions led to the practical construction by von Neumann of the first computer, followed by the computer revolution we are witnessing today.

The practical use of computers, and the unexpected depth of the concept of 'computation', has significantly expanded theoretical computer science, or TCS. In the last quarter century TCS has grown into a rich and beautiful field, making connections to other sciences and attracting first-rate young scientists. Here are just a few aspects of this evolution:

First, the focus of the field has changed from the notion of 'computation' to the much more elusive concept of 'efficient computation'. The fundamental notion of NP-completeness was formulated, its near-universal impact was gradually understood, and long-term goals, such as resolving the P vs NP question, were established. The theory of algorithms and a variety of computational models were developed. Randomness became a key tool and resource, revolutionizing the theory of algorithms.

Significantly, the emergence of the complexity-based notion of one-way function, together with the use of randomness, led to the development of modern cryptography. What many people at first thought were just mental games, such as trying to play poker without

cards, has turned into a powerful theory and practical systems of major economic importance. Complexity theory, which attempts to classify problems according to their computational difficulty, has integrated many of these ideas and has given rise to the field of proof complexity, where the goal is to quantify what constitutes a difficult proof.

Beyond these activities, which are internal to TCS, is important cross-fertilization between TCS and other subfields, such as combinatorics, algebra, topology and analysis. Moreover, the fundamental problems of TCS, notably 'P vs NP', have gained prominence as central problems of mathematics in general. More and more mathematicians are considering the computational aspects of their areas. In other words, they start with the fact 'An object exists' and follow it with the problem 'How fast can this object be found'?

A final aspect of TCS, which is to some people the most interesting, is that the field now overlaps with a whole new set of algorithmic problems from the sciences. In these problems the required output is not well defined in advance, and it may begin with almost any kind of data: a picture, a sonogram, readings from the Hubble Space Telescope, stock-market share values, DNA sequences, or neuron recordings of animals reacting to stimuli. Mathematical models are used to try to make sense of the data or predict their future values.

In general, the very notion of 'computation', and the major problems surrounding it, have taken on deep philosophical meaning and consequences. In addition to the P/NP question, the field is focused on a few clear and deep questions: for example, does randomization help computation? What constitutes a difficult theorem to prove? Can quantum mechanics be effectively simulated by classical means? The time is ripe for exciting growth and fundamental new understanding throughout this new field of theoretical computer science.

Quantum computing

Also related to the P/NP question is the investigation of quantum computing. This topic is closely related to P/NP because of a surprising demonstration in 1994 that if a quantum computer could be built, it would be capable of breaking any computer code now used or thought to be secure.

The need for a fundamentally new form of computing is very real – especially for running complex simulation models – and among several candidates, quantum computing may have the most promise. Although modern computers are extremely fast, they still use the classical binary calculating system of 0s and 1s that dates back 150 years to George Boole's adding machine. For many years this has seemed sufficient, especially in the presence of 'Moore's Law' – the observation that the capacity of computer chips doubles every two years or so while chip prices drop by half. This has been accomplished through better engineering and the production of smaller and smaller chips. But as we approach the millennium, we are also approaching quantum mechanical limits of chip size.

These limits were foreseen as early as 1982, when the American physicist Richard Feynman predicted that efforts to simulate quantum mechanical systems on digital computers would carry an inherent exponential 'overhead'. But in a lengthy side remark, he proposed that this difficulty might be circumvented by some form of quantum computer. In 1985, David Deutsch advanced the dialogue by suggesting that if quantum computers could be fast enough to solve quantum mechanical problems, they might also solve classical problems more quickly.

It appears that this is the case. In 1984, Peter W. Shor showed that a quantum computer could factor large numbers in time polynomial in the length of the numbers, a nearly exponential speed-up over classical algorithms. This was surprising for two reasons. First, modern cryptographers use long numbers as security codes because they are so difficult to factor – a job which a quantum computer could theoretically do rather quickly. Second, theoretical computer scientists had believed that no type of computing could be so much faster than conventional digital computers.

On the other hand, experimentalists are not at all sure they can build a quantum computer. As they pursue that possibility, there are numerous parallel efforts to design other kinds of computers based on principles other than Boolean arithmetic – all with the same goal of greatly expanding computer capacity. We can expect exciting and intense work in this field for years to come.

Maintaining the strength of mathematics in the 21st century

The bulk of this essay has been devoted to trends and problems of research. However, it is irresponsible to discuss research without mentioning the context in which it occurs. The success of research depends on the quality of the people doing it and the degree to which it receives sustained support from society; in other words, it requires 'patient capital'. The next millennium will bring a set of contextural questions every bit as challenging as the research we want to do.

Education

First, how can we attract the best young talent into mathematics? Here we have seen a significant change in the last half-century. During World War II, the systems

and techniques of science and technology generated much excitement, attracting post-war students to careers in research. This trend received a powerful stimulus in 1957, when the Soviet satellite *Sputnik* was launched and science was recognized for the political and economic power it could generate. Research became as important to society as it was fascinating to practitioners.

Towards the end of the century, however, society's interest in many areas of research appears to have diminished in both developed and developing countries. Many bright students who would once have chosen careers in mathematics or science are not doing so; they are choosing applied information science, business, or other areas where the future looks more interesting. Certainly there are practical reasons for bypassing mathematics, such as the difficulty of the subject, the long period of study required for entry, and the comparatively modest salaries. But there appears also to be a fundamental lack of appreciation for the richness and relevance of the subject itself.

An obvious reason for student disinterest is that we are not communicating a complete picture of mathematics as a field where one may choose among many intellectually rewarding and challenging careers. The people best positioned to do this communication are high school teachers, college professors, and fellow students. However, these groups can only describe current opportunities and fast-growing fields if they, in turn, are informed by those in the profession. The mathematics community as a whole can help by fostering more interaction at every level of teaching and practice, and by widening channels of communication with the students who will eventually replace us and extend our work into the next century.

It is ironic that student interest is low at a time when career opportunities for professional mathematicians have never been greater or more diverse. This is true both for the traditional disciplinary areas, which are rich with new developments and challenging problems, and in the more applied fields and other areas of science, where demand for mathematicians with proper training will continue to grow rapidly in the foreseeable future.

Outreach

Closely allied with educational needs is the opportunity to better communicate with public about mathematical issues. Mathematicians understand the purpose and value of their work, but many people in government, business, and even in education do not. If mathematicians expect their research at universities to be supported by public funds, which is customary, we must present a vivid picture of that research and its power. It is no longer acceptable to remain aloof from the pressing needs of the world or to work in an ivory tower.

Interactivity

Finally, the trend towards interactivity merits a final mention. We have seen that within mathematics and throughout the sciences, much of the most productive work is being done at frontiers between subfields, fields, and disciplines. Mathematics loses something when it is isolated or fragmented by disciplinary paradigms. However, many institutions have been slow to adapt to this reality. Universities around the world, and many industries and government agencies, stand to gain much by removing barriers to collaborations.

In particular, much can be done to enhance interactions between academic and industrial mathematicians. The primary missions of academia and industry are different, but the two cultures have much to gain from collaboration. In general, the scientific enterprise can function at full potential only when there is a fast flow of knowledge between the creators and users of mathematics.

The next millennium

A new and powerful trend that will carry us far into the next millennium is characterized by the globalization, interactivity, and 'opening out' of the mathematical enterprise. As a harbinger of things to come, we note the mode in which Thomas Hales chose to announce his proof of the Kepler sphere-packing problem. Rather than publish it in a journal, where his results would be seen by a small number of specialists, he opened it to an unlimited audience via the Internet. In addition, he frankly invited scrutiny of his proof and further contributions to its accuracy – a significant step in the competitive world of top-level mathematics.

In general, then, we mathematicians have two objectives as we enter the next millennium. The first is to maintain traditional strengths in basic research, which is the seedbed of new thinking and new applications. Second, we are called to broaden our exploration of the terrain outside the traditional boundaries of our field – to the other sciences and to the world beyond science. With each passing year, mathematicians achieve more effectiveness in their work to the degree that they offer it to others and include others in the world of mathematics.

Received 2 July 1999; accepted 9 July 1999