THE GREAT MATHEMATICAL PROBLEMS

Marvels and Mysteries of Mathematics

IAN STEWART
Great problems

TELEVISION PROGRAMMES ABOUT MATHEMATICS are rare, good ones rarer. One of the best, in terms of audience involvement and interest as well as content, was Fermat’s last theorem. The programme was produced by John Lynch for the British Broadcasting Corporation’s flagship popular science series Horizon in 1996. Simon Singh, who was also involved in its making, turned the story into a spectacular bestselling book. On a website, he pointed out that the programme’s stunning success was a surprise:

It was 50 minutes of mathematicians talking about mathematics, which is not the obvious recipe for a TV blockbuster, but the result was a programme that captured the public imagination and which received critical acclaim. The programme won the BAFTA for best documentary, a Priz Italia, other international prizes and an Emmy nomination – this proves that mathematics can be as emotional and as gripping as any other subject on the planet.

I think that there are several reasons for the success of both the television programme and the book and they have implications for the stories I want to tell here. To keep the discussion focused, I’ll concentrate on the television documentary.

Fermat’s last theorem is one of the truly great mathematical problems, arising from an apparently innocuous remark which one of the leading mathematicians of the seventeenth century wrote in the margin of a classic textbook. The problem became notorious because no one could prove what Pierre de Fermat’s marginal note claimed, and
this state of affairs continued for more than 300 years despite strenuous efforts by extraordinarily clever people. So when the British mathematician Andrew Wiles finally cracked the problem in 1995, the magnitude of his achievement was obvious to anyone. You didn’t even need to know what the problem was, let alone how he had solved it. It was the mathematical equivalent of the first ascent of Mount Everest.

In addition to its significance for mathematics, Wiles’s solution also involved a massive human-interest story. At the age of ten, he had become so intrigued by the problem that he decided to become a mathematician and solve it. He carried out the first part of the plan, and got as far as specialising in number theory, the general area to which Fermat’s last theorem belongs. But the more he learned about real mathematics, the more impossible the whole enterprise seemed. Fermat’s last theorem was a baffling curiosity, an isolated question of the kind that any number theorist could dream up without a shred of convincing evidence. It didn’t fit into any powerful body of technique. In a letter to Heinrich Olbers, the great Gauss had dismissed it out of hand, saying that the problem had ‘little interest for me, since a multitude of such propositions, which one can neither prove nor refute, can easily be formulated’. Wiles decided that his childhood dream had been unrealistic and put Fermat on the back burner. But then, miraculously, other mathematicians suddenly made a breakthrough that linked the problem to a core topic in number theory, one on which Wiles was already an expert. Gauss, uncharacteristically, had underestimated the problem’s significance, and was unaware that it could be linked to a deep, though apparently unrelated, area of mathematics.

With this link established, Wiles could now work on Fermat’s enigma and do credible research in modern number theory at the same time. Better still, if Fermat didn’t work out, anything significant that he discovered while trying to prove it would be publishable in its own right. So off the back burner it came, and Wiles began to think about Fermat’s problem in earnest. After seven years of obsessive research, carried on in private and in secret – an unusual precaution in mathematics – he became convinced that he had found a solution. He delivered a series of lectures at a prestigious number theory conference, under an obscure title that fooled no one. The exciting news broke, in
the media as well as the halls of academe: Fermat’s last theorem had been proved.

The proof was impressive and elegant, full of good ideas. Unfortunately, experts quickly discovered a serious gap in its logic. In attempts to demolish great unsolved problems of mathematics, this kind of development is depressingly common, and it almost always proves fatal. However, for once the Fates were kind. With assistance from his former student Richard Taylor, Wiles managed to bridge the gap, repair the proof, and complete his solution. The emotional burden involved became vividly clear in the television programme: it must have been the only occasion when a mathematician has burst into tears on screen, just recalling the traumatic events and the eventual triumph.

You may have noticed that I haven’t told you what Fermat’s last theorem is. That’s deliberate; it will be dealt with in its proper place. As far as the success of the television programme goes, it doesn’t actually matter. In fact, mathematicians have never greatly cared whether the theorem that Fermat scribbled in his margin is true or false, because nothing of great import hangs on the answer. So why all the fuss? Because a huge amount hangs on the inability of the mathematical community to find the answer. It’s not just a blow to our self-esteem: it means that existing mathematical theories are missing something vital. In addition, the theorem is very easy to state; this adds to its air of mystery. How can something that seems so simple turn out to be so hard?

Although mathematicians didn’t really care about the answer, they cared deeply that they didn’t know what it was. And they cared even more about finding a method that could solve it, because that must surely shed light not just on Fermat’s question, but on a host of others. This is often the case with great mathematical problems: it is the methods used to solve them, rather than the results themselves, that matter most. Of course, sometimes the actual result matters too: it depends on what its consequences are.

Wiles’s solution is much too complicated and technical for television; in fact, the details are accessible only to specialists. The proof does involve a nice mathematical story, as we’ll see in due course, but any attempt to explain that on television would have lost most of the audience immediately. Instead, the programme sensibly
concentrated on a more personal question: what is it like to tackle a notoriously difficult mathematical problem that carries a lot of historical baggage? Viewers were shown that there existed a small but dedicated band of mathematicians, scattered across the globe, who cared deeply about their research area, talked to each other, took note of each other’s work, and devoted a large part of their lives to advancing mathematical knowledge. Their emotional investment and social interaction came over vividly. These were not clever automata, but real people, engaged with their subject. That was the message.

Those are three big reasons why the programme was such a success: a major problem, a hero with a wonderful human story, and a supporting cast of emotionally involved people. But I suspect there was a fourth, not quite so worthy. The majority of non-mathematicians seldom hear about new developments in the subject, for a variety of perfectly sensible reasons: they’re not terribly interested anyway; newspapers hardly ever mention anything mathematical; when they do, it’s often facetious or trivial; and nothing much in daily life seems to be affected by whatever it is that mathematicians are doing behind the scenes. All too often, school mathematics is presented as a closed book in which every question has an answer. Students can easily come to imagine that new mathematics is as rare as hen’s teeth.

From this point of view, the big news was not that Fermat’s last theorem had been proved. It was that at last someone had done some new mathematics. Since it had taken mathematicians more than 300 years to find a solution, many viewers subconsciously concluded that the breakthrough was the first important new mathematics discovered in the last 300 years. I’m not suggesting that they explicitly believed that. It ceases to be a sustainable position as soon as you ask some obvious questions, such as ‘Why does the Government spend good money on university mathematics departments?’ But subconsciously it was a common default assumption, unquestioned and unexamined. It made the magnitude of Wiles’s achievement seem even greater.

One of the aims of this book is to show that mathematical research is thriving, with new discoveries being made all the time. You don’t hear much about this activity because most of it is too technical for non-specialists, because most of the media are wary of anything intellectually more challenging than The X Factor, and because the applications of mathematics are deliberately hidden away to avoid
causing alarm. ‘What? My iPhone depends on advanced mathematics? How will I log in to Facebook when I failed my maths exams?’

Historically, new mathematics often arises from discoveries in other areas. When Isaac Newton worked out his laws of motion and his law of gravity, which together describe the motion of the planets, he did not polish off the problem of understanding the solar system. On the contrary, mathematicians had to grapple with a whole new range of questions: yes, we know the laws, but what do they imply? Newton invented calculus to answer that question, but his new method also has limitations. Often it rephrases the question instead of providing the answer. It turns the problem into a special kind of formula, called a differential equation, whose solution is the answer. But you still have to solve the equation. Nevertheless, calculus was a brilliant start. It showed us that answers were possible, and it provided one effective way to seek them, which continues to provide major insights more than 300 years later.

As humanity’s collective mathematical knowledge grew, a second source of inspiration started to play an increasing role in the creation of even more: the internal demands of mathematics itself. If, for example, you know how to solve algebraic equations of the first, second, third, and fourth degree, then you don’t need much imagination to ask about the fifth degree. (The degree is basically a measure of complexity, but you don’t even need to know what it is to ask the obvious question.) If a solution proves elusive, as it did, that fact alone makes mathematicians even more determined to find an answer, whether or not the result has useful applications. I’m not suggesting applications don’t matter. But if a particular piece of mathematics keeps appearing in questions about the physics of waves – ocean waves, vibrations, sound, light – then it surely makes sense to investigate the gadget concerned in its own right. You don’t need to know ahead of time exactly how any new idea will be used: the topic of waves is common to so many important areas that significant new insights are bound to be useful for something. In this case, those somethings included radio, television, and radar. If somebody thinks up a new way to understand heat flow, and comes up with a brilliant new technique that unfortunately lacks proper mathematical support,
then it makes sense to sort the whole thing out as a piece of mathematics. Even if you don’t give a fig about how heat flows, the results might well be applicable elsewhere. Fourier analysis, which emerged from this particular line of investigation, is arguably the most useful single mathematical idea ever found. It underpins modern telecommunications, makes digital cameras possible, helps to clean up old movies and recordings, and a modern extension is used by the FBI to store fingerprint records.7

After a few thousand years of this kind of interchange between the external uses of mathematics and its internal structure, these two aspects of the subject have become so densely interwoven that picking them apart is almost impossible. The mental attitudes involved are more readily distinguishable, though, leading to a broad classification of mathematics into two kinds: pure and applied. This is defensible as a rough-and-ready way to locate mathematical ideas in the intellectual landscape, but it’s not a terribly accurate description of the subject itself. At best it distinguishes two ends of a continuous spectrum of mathematical styles. At worst, it misrepresents which parts of the subject are useful and where the ideas come from. As with all branches of science, what gives mathematics its power is the combination of abstract reasoning and inspiration from the outside world, each feeding off the other. Not only is it impossible to pick the two strands apart: it’s pointless.

Most of the really important mathematical problems, the great problems that this book is about, have arisen within the subject through a kind of intellectual navel-gazing. The reason is simple: they are mathematical problems. Mathematics often looks like a collection of isolated areas, each with its own special techniques: algebra, geometry, trigonometry, analysis, combinatorics, probability. It tends to be taught that way, with good reason: locating each separate topic in a single well-defined area helps students to organise the material in their minds. It’s a reasonable first approximation to the structure of mathematics, especially long-established mathematics. At the research frontiers, however, this tidy delineation often breaks down. It’s not just that the boundaries between the major areas of mathematics are blurred. It’s that they don’t really exist.

Every research mathematician is aware that, at any moment, suddenly and unpredictably, the problem they are working on may turn
out to require ideas from some apparently unrelated area. Indeed, new research often combines areas. For instance, my own research mostly centres on pattern formation in dynamical systems, systems that change over time according to specific rules. A typical example is the way animals move. A trotting horse repeats the same sequence of leg movements over and over again, and there is a clear pattern: the legs hit the ground together in diagonally related pairs. That is, first the front left and back right legs hit, then the other two. Is this a problem about patterns, in which case the appropriate methods come from group theory, the algebra of symmetry? Or is it a problem about dynamics, in which case the appropriate area is Newtonian-style differential equations?

The answer is that, by definition, it has to be both. It is not their intersection, which would be the material they have in common—basically, nothing. Instead, it is a new ‘area’, which straddles two of the traditional divisions of mathematics. It is like a bridge across a river that separates two countries; it links the two, but belongs to neither. But this bridge is not a thin strip of roadway; it is comparable in size to each of the countries. Even more vitally, the methods involved are not limited to those two areas. In fact, virtually every course in mathematics that I have ever studied has played a role somewhere in my research. My Galois theory course as an undergraduate at Cambridge was about how to solve (more precisely, why we can’t solve) an algebraic equation of the fifth degree. My graph theory course was about networks, dots joined by lines. I never took a course in dynamical systems, because my PhD was in algebra, but over the years I picked up the basics, from steady states to chaos. Galois theory, graph theory, dynamical systems: three separate areas. Or so I assumed until 2011, when I wanted to understand how to detect chaotic dynamics in a network of dynamical systems, and a crucial step depended on things I’d learned 45 years earlier in my Galois theory course.

Mathematics, then, is not like a political map of the world, with each speciality neatly surrounded by a clear boundary, each country tidily distinguished from its neighbours by being coloured pink, green, or pale blue. It is more like a natural landscape, where you can never really say where the valley ends and the foothills begin, where the forest merges into woodland, scrub, and grassy plains, where lakes insert
regions of water into every other kind of terrain, where rivers link the snow-clad slopes of the mountains to the distant, low-lying oceans. But this ever-changing mathematical landscape consists not of rocks, water, and plants, but of ideas; it is tied together not by geography, but by logic. And it is a dynamic landscape, which changes as new ideas and methods are discovered or invented. Important concepts with extensive implications are like mountain peaks, techniques with lots of uses are like broad rivers that carry travellers across the fertile plains. The more clearly defined the landscape becomes, the easier it is to spot unscaled peaks, or unexplored terrain that creates unwanted obstacles. Over time, some of the peaks and obstacles acquire iconic status. These are the great problems.

What makes a great mathematical problem great? Intellectual depth, combined with simplicity and elegance. Plus: it has to be hard. Anyone can climb a hillock; Everest is another matter entirely. A great problem is usually simple to state, although the terms required may be elementary or highly technical. The statements of Fermat’s last theorem and the four colour problem make immediate sense to anyone familiar with school mathematics. In contrast, it is impossible even to state the Hodge conjecture or the mass gap hypothesis without invoking deep concepts at the research frontiers – the latter, after all, comes from quantum field theory. However, to those versed in such areas, the statement of the question concerned is simple and natural. It does not involve pages and pages of dense, impenetrable text. In between are problems that require something at the level of undergraduate mathematics, if you want to understand them in complete detail. A more general feeling for the essentials of the problem – where it came from, why it’s important, what you could do if you possessed a solution – is usually accessible to any interested person, and that’s what I will be attempting to provide. I admit that the Hodge conjecture is a hard nut to crack in that respect, because it is very technical and very abstract. However, it is one of the seven Clay Institute millennium mathematics problems, with a million-dollar prize attached, and it absolutely must be included.

Great problems are creative: they help to bring new mathematics into being. In 1900 David Hilbert delivered a lecture at the
International Congress of Mathematicians in Paris, in which he listed 23 of the most important problems in mathematics. He didn’t include Fermat’s last theorem, but he mentioned it in his introduction. When a distinguished mathematician lists what he thinks are some of the great problems, other mathematicians pay attention. The problems wouldn’t be on the list unless they were important, and hard. It is natural to rise to the challenge, and try to answer them. Ever since, solving one of Hilbert’s problems has been a good way to win your mathematical spurs. Many of these problems are too technical to include here, many are open-ended programmes rather than specific problems, and several appear later in their own right. But they deserve to be mentioned, so I’ve put a brief summary in the notes. That’s what makes a great mathematical problem great. What makes it problematic is seldom deciding what the answer should be. For virtually all great problems, mathematicians have a very clear idea of what the answer ought to be — or had one, if a solution is now known. Indeed, the statement of the problem often includes the expected answer. Anything described as a conjecture is like that: a plausible guess, based on a variety of evidence. Most well-studied conjectures eventually turn out to be correct, though not all. Older terms like hypothesis carry the same meaning, and in the Fermat case the word ‘theorem’ is (more precisely, was) abused — a theorem requires a proof, but that was precisely what was missing until Wiles came along.

Proof, in fact, is the requirement that makes great problems problematic. Anyone moderately competent can carry out a few calculations, spot an apparent pattern, and distil its essence into a pithy statement. Mathematicians demand more evidence than that: they insist on a complete, logically impeccable proof. Or, if the answer turns out to be negative, a disproof. It isn’t really possible to appreciate the seductive allure of a great problem without appreciating the vital role of proof in the mathematical enterprise. Anyone can make an educated guess. What’s hard is to prove it’s right. Or wrong.

The concept of mathematical proof has changed over the course of history, with the logical requirements generally becoming more stringent. There have been many highbrow philosophical discussions of the nature of proof, and these have raised some important issues. Precise logical definitions of ‘proof’ have been proposed and
implemented. The one we teach to undergraduates is that a proof begins with a collection of explicit assumptions called axioms. The axioms are, so to speak, the rules of the game. Other axioms are possible, but they lead to different games. It was Euclid, the ancient Greek geometer, who introduced this approach to mathematics, and it is still valid today. Having agreed on the axioms, a proof of some statement is a series of steps, each of which is a logical consequence of either the axioms, or previously proved statements, or both. In effect, the mathematician is exploring a logical maze, whose junctions are statements and whose passages are valid deductions. A proof is a path through the maze, starting from the axioms. What it proves is the statement at which it terminates.

However, this tidy concept of proof is not the whole story. It’s not even the most important part of the story. It’s like saying that a symphony is a sequence of musical notes, subject to the rules of harmony. It misses out all of the creativity. It doesn’t tell us how to find proofs, or even how to validate other people’s proofs. It doesn’t tell us which locations in the maze are significant. It doesn’t tell us which paths are elegant and which are ugly, which are important and which are irrelevant. It is a formal, mechanical description of a process that has many other aspects, notably a human dimension. Proofs are discovered by people, and research in mathematics is not just a matter of step-by-step logic.

Taking the formal definition of proof literally can lead to proofs that are virtually unreadable, because most of the time is spent dotting logical i’s and crossing logical t’s in circumstances where the outcome already stares you in the face. So practising mathematicians cut to the chase, and leave out anything that is routine or obvious. They make it clear that there’s a gap by using stock phrases like ‘it is easy to verify that’ or ‘routine calculations imply’. What they don’t do, at least not consciously, is to slither past a logical difficulty and to try to pretend it’s not there. In fact, a competent mathematician will go out of his or her way to point out exactly those parts of the argument that are logically fragile, and they will devote most of their time to explaining how to make them sufficiently robust. The upshot is that a proof, in practice, is a mathematical story with its own narrative flow. It has a beginning, a middle, and an end. It often has subplots, growing out of the main plot, each with its own resolution. The British mathematician