Introduction to Fermat's Last Theorem

David A. Cox*

The announcement last summer of a proof of Fermat's Last Theorem was an exciting event for the entire mathematics community. This article will discuss the mathematical history of Fermat's Last Theorem (which we will abbreviate throughout as FLT), broken up into the following periods:

1. Diophantus to Euler (250–1783 A.D.)
2. Euler to Frey (1783–1982 A.D.)
3. Frey to Wiles (1982–1993 A.D.)

We will give only an introduction to the story of Fermat's Last Theorem, and our account is by no means definitive. Many of the more technical terms are not defined completely and the few proofs that appear are only sketched. On the other hand, I hope that the article succeeds in conveying the flavor of this truly wonderful mathematics.

1. DIOPHANTUS TO EULER. Our history of FLT starts around 250 A.D. with Diophantus, whose *Arithmetica* considered many problems in elementary number theory. Consider Problem 8 from Book II, which asks "to divide a given square number into two squares" ([**10**], p. 144). Diophantus' solution is as follows: Let the given square be 16, let x^2 be one of the required squares and $(2x - 4)^2$ the other square. Therefore, we must satisfy $x^2 + (2x - 4)^2 = 16$, which implies

$$x^2 + 4x^2 - 16x + 16 = 16 \Rightarrow 5x^2 = 16x \Rightarrow x = 16/5.$$

Hence the required squares are $256/25$ and $144/25$.

We can observe two things about this problem. First, solutions are presumed to be rational. We neither restrict to only integer solutions nor generalize to real solutions. Second, we care only about finding one solution to a given problem; if we find one, we are happy and move on.

The *Arithmetica* was one of the last Greek mathematical works translated into Latin; this occured in 1575. Fermat (1601–1665) had a copy of Bachet's translation of 1621 and made a series of intriguing annotations in its margins. Sometime in the late 1630's, while thinking about the problem given above, he added the famous words in the margin:

> "On the other hand, it is impossible to separate a cube into two cubes, or a biquadrate into two biquadrates, or generally any power except a square into

*This article is based on a lecture given at the 1993 Smith College Regional Geometry Institute. This audience included high schools teachers, undergraduates, graduate students and researchers in discrete and computational geometry. I would like to thank Thomas Colthurst for transcribing the lecture, and I am grateful to my colleagues who pointed out errors in earlier versions of the manuscript.

two powers with the same exponent. I have discovered a truly marvellous proof of this, which however the margin is not large enough to contain." ([**10**], pp. 144–145)

Hence the basic claim of Fermat's Last Theorem is that the equation $x^n + y^n = z^n$ has no solutions when x, y, z are nonzero integers and $n > 2$. Generations of mathematical historians have debated over whether Fermat really did have a proof, though many experts doubt that he did. For one thing, the equation $x^n + y^n = z^n$ was atypical for Fermat—the vast majority of the other equations he studied dealt with exponents ≤ 4. Also, in his correspondence, he only stated FLT for the exponent $n = 3$. As for Fermat's "marvellous proof," it probably used the technique of infinite descent. His descent argument for $n = 4$ is actually known: it can be found in Fermat's proof that the area of a right triangle with integral sides cannot be a square. This proof is given in one of his marginal notes, although even here, Fermat complains that there isn't enough room to give the proof "with all detail" ([**10**], p. 293). It seems likely that Fermat thought that his proofs for $n = 3$ and 4 generalized, and they almost certainly didn't.

So, what happened after Fermat? In 1670, his marginal notes were published by his son, and some of his letters appeared in Wallis' *Opera Mathematica*. In 1729, Goldbach wrote Euler and mentioned some of Fermat's results. This got Euler, only 22 at the time, thinking about number theory. Three years later, Euler wrote his first paper on number theory, disproving a conjecture of Fermat's on primes of the form $2^{2^n} + 1$. For the next fifty years, Euler proved many of Fermat's conjectures and in so doing, transformed number theory from a collection of miscellaneous facts and results into an organized field at the very center of mathematics.

Here is an example of what Euler did. In Problem 17 of Book VI of the *Arithmetica* (Problem 19 in Bachet's numbering), Fermat had written in the margin, "Can one find in whole numbers a square different from 25, when increased by 2, becomes a cube? ... [The answer involves] the doctrine of whole numbers, which is assuredly very beautiful and very subtle ..." ([**6**], p. 269). In modern terms, Fermat is claiming that the only integer solutions to $x^3 = y^2 + 2$ are given by $(x, y) = (3, \pm 5)$. You can see how the emphasis is different from Diophantus—Fermat is looking for *all* solutions, and he recognizes that asking for integer solutions (rather than rational ones) is a question of independent interest.

To prove Fermat's claim, Euler ([**4**], Part II, §193) uses numbers of the form $a + b\sqrt{-2}$, with a, b integers. First observe

$$x^3 = y^2 + 2 = \left(y + \sqrt{-2}\right)\left(y - \sqrt{-2}\right).$$

One can show that $y + \sqrt{-2}$ and $y - \sqrt{-2}$ are relatively prime,and since their product is a cube, each of them must also be a cube. Thus there is a number $p + q\sqrt{-2}$ such that

$$y + \sqrt{-2} = \left(p + q\sqrt{-2}\right)^3 = p^3 - 6pq^2 + \left(3p^3q - 2q^3\right)\sqrt{-2}$$
$$\Rightarrow 1 = 3p^2q - 2q^3 = q\left(3p^2 - 2q^2\right).$$

The last equation implies $p = \pm 1$ and $q = 1$. Substituting this in, we get $y = p^3 - 6pq^2 = \pm 5$ and $x = 3$, as claimed.

This proof, while elegant, is incomplete, for we do not know that numbers of the form $a + b\sqrt{-2}$ have *unique factorization*, or even for that matter, *primes* (although it is relatively easy to prove that the numbers $a + b\sqrt{-2}$ have these

properties). There are several reasons why this example is important:

- First, it reminds us that there are lots of diophantine equations besides just FLT, and what we really want is a method for dealing with as many of them as possible.
- Second, it shows that properties of integers (such as unique factorization) can apply in more general situations, and it illustrates how a result in one context (the integers) can be proved by working in a more general context (numbers of the form $a + b\sqrt{-2}$).
- Finally, the equation $y^2 = x^3 - 2$ is an example of an *elliptic curve*. Elliptic curves will play a crucial role in the final proof of FLT.

2. EULER TO FREY. This section is only a sketch of more than two hundred years of beautiful and wonderful number theory. For more information on the work on FLT done during this period, we warmly recommend both Edwards' *Fermat's Last Theorem* [3] and Ribenboim's *13 Lectures on Fermat's Last Theorem* [21]. (Precise references for results mentioned in this section can be found in these books.)

Before we begin, first observe that it suffices to prove FLT for $n = 4$ (done by Fermat) and for n an odd prime (since we can factor the exponent). We can also assume that x, y, z are nonzero relatively prime integers (because we can cancel common factors). That being said, here are some of the highlights of the 19th century work on FLT:

- By the early 1800s, all of Fermat's problems were solved except for FLT (thus justifying the name, Fermat's Last Theorem).
- 1816—The French Academy announces a prize for a solution to FLT.
- In the 1820's Sophie Germain shows that if p and $2p + 1$ are prime, then $x^p + y^p = z^p$ has no solution with $p \nmid xyz$. This is the so-called Case I of FLT. (Case II is where $p|xyz$ and is usually regarded as being much harder.)
- 1825—Dirichlet and Legendre prove FLT for $n = 5$.
- 1832—Dirichlet, after trying to prove it for $n = 7$, proves FLT for $n = 14$.
- 1839—Lamé proves FLT for $n = 7$.
- 1847—Lamé and Cauchy present false proofs of FLT for general n.
- 1844–1847—Kummer's work on FLT.

Let us describe Kummer's work on FLT in more detail. Kummer (and Cauchy and Lamé) started, à la Euler, by factoring the right hand side of the FLT equation as

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

where $\zeta = e^{2\pi i/p} = \cos(2\pi/p) + i \sin(2\pi/p)$ is a *pth root of unity* and satisfies $\zeta^p = 1$. In general, working with roots of unity will require us to use numbers of the form

$$a_0 + a_1\zeta + \cdots + a_{p-1}\zeta^{p-1}, a_0, \ldots, a_{p-1} \in \mathbf{Z},$$

which are called *cyclotomic integers*. But a problem arises when unique factorization, one of our main tools, fails for the cyclotomic integers. As Kummer discovered in 1844, this first occurs for $p = 23$ (and in fact, unique factorization fails for all bigger primes as well).

Kummer's solution to this was twofold. First, he introduced a generalization of cyclotomic integers, called *ideal numbers*, which make up for the lack of unique

factorization. Second, he defined the *class number* h, which measures how badly unique factorization fails.

Here is a summary of Kummer's results:

- 1847—Theorem: FLT holds for p if $p \nmid h$ (such p are called *regular primes*).
- 1847—Theorem: p is regular iff p doesn't divide the numerator of the Bernoulli numbers $B_2, B_4, \ldots, B_{p-3}$.
  We can define the Bernoulli numbers by the power series

$$\frac{x}{e^x - 1} = \sum_{n=1}^{\infty} \frac{B_n}{n!} x^n.$$

  A corollary of this result is that for $p < 100$, only 37, 59 and 67 are irregular.
- 1850—The French Academy offers a second prize for a solution to FLT.
- 1856—At Cauchy's suggestion, the Academy withdraws the prize and then awards a medal to Kummer.
- 1857—Kummer develops complicated criteria for proving FLT for certain irregular primes. There are some gaps in his proofs which are later filled in by Vandiver in the 1920s. These results establish FLT for $p < 100$.

The above history makes a wonderful story about how FLT inspired one of the greatest inventions in number theory, but the story is unfortunately false. Kummer was actually not trying to prove FLT, but something called a reciprocity theorem. Reciprocity theorems have their origins in Fermat's study of equations like $p = x^2 + y^2$ and $p = x^2 + 2y^2$, where p is a prime. In trying to understand these results, Euler, Lagrange, Legendre and Gauss created the theory of quadratic forms and proved the law of quadratic reciprocity. Later, Gauss, Abel and Jacobi formulated versions of cubic and biquadratic reciprocity, and Kummer and Eisenstein made the first attempts at higher reciprocity laws. Cyclotomic integers and ideal numbers came about primarily from Kummer's attempts to prove these higher reciprocity laws. In turn, these concepts not only had something interesting to say about FLT, but they also made significant contributions toward the development of class field theory and abstract algebra (we use the terminology "ideal of a ring" because of Kummer's "ideal numbers").

Here are some highlights of the history of FLT after Kummer:

- 1908—The Wolfskehl prize for a solution to FLT is announced. Later inflation in the German mark reduces the value of this prize considerably, but does not reduce the flow of crank solutions submitted.
- 1909—Wieferich proves if $x^p + y^p = z^p$ and $p \nmid xyz$ (Case I of FLT), then $2^{p-1} \equiv 1 \bmod p^2$. This is a strong congruence which is particularly easy to check on a computer.
- 1953—Inkeri proves that if $x^p + y^p = z^p$ and $x < y < z$, then $x > ((2p^3 + p)/\log(3p))^p$ in Case I and $x > p^{3p-4}$ in Case II.
- 1971—Brillhart, Tonascia and Weinberger show that Case I of FLT is true for all primes less than $3 \cdot 10^9$.
- 1976—Wagstaff shows that FLT is true for all primes less than 125,000.

These results imply that any counterexample to FLT must involve $p \geq 125{,}003$ and $z > y > x > (125{,}003)^{375{,}005} \approx 4.5 \cdot 10^{1{,}911{,}370}$. (In 1992, as a byproduct of other computations, the lower bound on the exponent was raised to $p > 4{,}000{,}000$ —see [2].)

We should also mention that the Fermat equation $x^n + y^n = z^n$ has been studied in many other contexts, including polynomials, entire functions and matrices (see [21] and, for a recent proof of the polynomial case, [18]).

3. FREY TO WILES. In 1983, Faltings [4] proved the Mordell Conjecture, which implies that a polynomial equation with rational coefficients $Q(x, y) = 0$ has only finitely many rational solutions when the curve has genus ≥ 2 (for a definition of genus, see the sidebar "The genus of an algebraic curve"). Since $x^n + y^n = 1$ has genus ≥ 2 for $n \geq 4$, there are only finitely many rational solutions by the Mordell Conjecture. Then, clearing denominators, it follows easily that $x^n + y^n = z^n$ has only finitely many relatively prime integer solutions.

The genus of an algebraic curve

The *genus* of a curve given by a polynomial equation $p(x, y) = 0$ of degree n can be defined in a variety of ways. When the equation is sufficiently smooth (which is true for the Fermat curve $x^n + y^n = 1$), then the genus is $g = (n - 1)(n - 2)/2$. This is ≥ 2 when $n \geq 4$.

Topologically, the solutions of $p(x, y) = 0$ over the complex numbers form a compact Riemann surface minus a finite set of points, and then the genus is just the usual genus of this compact real 2-dimensional manifold.

Analytically, a Riemann surface is a compact complex 1-dimensional manifold, and one can define the notion of a *holomorphic 1-form*. Then the genus is the maximum number of linearly independent holomorphic 1-forms on the surface.

For example, the Riemann sphere has genus zero, so that there are no holomorphic 1-forms, while the elliptic curve $y^2 = Ax^3 + Bx^2 + Cx + D$ has genus 1, and up to a constant, dx/y is the only holomorphic 1-form.

This may not seem so useful, since we want to show that the number of solutions is actually zero. But Granville [9] and Heath-Brown [11], aided by an observation of Filaseta, used the above finiteness result to show that FLT holds for "most" exponents, in the sense that if you look at all exponents—prime and composite—from 3 to n, the percentage where FLT could fail approaches zero as n increases (see [28] for the details). Also, Adelman and Heath-Brown [1] showed that Case I of FLT was true for infinitely many prime exponents.

For us, the Mordell Conjecture is interesting because it shows how a general conjecture in number theory can have some consequences concerning FLT. Also, in proving the Mordell Conjecture, Faltings used the machinery of modern algebraic geometry, which had been developing since the 1950's.

By the end of the 1980's, there were several conjectures in number theory which, if proved, would imply FLT, though sometimes only for sufficiently large exponents (see the sidebar "Conjectures that imply Fermat's Last Theorem"). This showed that FLT was not an isolated oddity, but rather was intimately connected to other parts of number theory. People were especially excited in 1988, when Miyaoka gave a lecture in Bonn in which he stated one of these conjectures, the arithmetic Bogomolov-Miyaoka-Yau inequality, as a theorem. This would have proved FLT for all large primes p (without saying explicitly what "large" meant). In the days following his lecture, there was much fanfare in the press, and it was rather disappointing when a week later an error was found in the argument.

Conjectures that imply Fermat's Last Theorem

By the late 1980's, there were several conjectures in number theory which, if proved, would imply FLT, at least for large exponents:

Diophantine Geometry. The following conjectures in diophantine geometry would imply FLT for all sufficiently large exponents:

- The *abc* Conjecture states that if a, b, c are relatively prime integers with $a + b = c$, then $\max(|a|, |b|, |c|)$ is bounded in terms of the primes dividing abc.
- Szpiro's Conjecture relates the minimal discriminant to the conductor of an elliptic curve. These terms are discussed in the sidebar entitled "Invariants of the Frey curve".
- Vojta's Conjecture concerns heights of points (relative to the canonical class) of a curve defined over the integers.

Precise statements of these conjectures (and the relations among them) can be found in Lang's survey article **[18]**.

Arithmetic Surfaces. The Bogomolov-Miyaoka-Yau inequality for arithmetic surfaces relates various invariants of a curve defined over the integers. This inequality is an arithmetic analog of a well known inequality for complex surfaces. By Parshin **[20]** and Vojta (Appendix to **[16]**), this conjecture implies versions of the above diophantine conjectures strong enough to prove FLT for all large exponents.

Elliptic Curves. The Taniyama-Shimura Conjecture states that all elliptic curves over the rational numbers are modular (a more precise statement of the conjecture is in the body of the article). As we will explain, the work of Frey, Serre and Ribet shows that this conjecture implies FLT for *all* exponents.

Of these conjectures, the one ultimately most important for FLT is the Taniyama-Shimura Conjecture, which asserts that all elliptic curves over $\mathbf{Q}$ are modular (this term will be defined below). The full story of this conjecture goes back to Jacobi and Riemann, but we will begin our account with the work of Gerhard Frey from 1982 to 1986 (see **[7]** and **[8]**). Frey showed that nontrivial solutions to FLT give rise to very special elliptic curves, which we shall call *Frey curves*. His basic insight was that Frey curves were so special that they couldn't be modular. Hence, if the Taniyama-Shimura Conjecture were true, Frey curves couldn't exist, and FLT would follow.

If $a^p + b^p = c^p$ is a solution to FLT, then the associated Frey curve is

$$y^2 = x(x - a^p)(x + b^p).$$

As usual, we assume a, b, c are nonzero relatively prime integers and p is an odd prime. This is an elliptic curve over the rational numbers $\mathbf{Q}$, similar to the equation $y^2 = x^3 - 2$ considered by Fermat. In general, an *elliptic curve over* $\mathbf{Q}$ is given by an equation of the form

$$y^2 = Ax^3 + Bx^2 + Cx + D,$$

where A, B, C, D are rational and the cubic polynomial in x on the right hand side of the equation has distinct roots. Elliptic curves are a large and important part of modern number theory.

Actually, we have to be a bit careful when constructing the Frey curve. A solution $a^p + b^p = c^p$ gives rise to solutions $b^p + a^p = c^p$ and $a^p + (-c)^p = (-b)^p$ (since p is odd). From here it is easy to rearrange the solution so that b is even and $a \equiv -1 \bmod 4$. This is needed in order that the Frey curve be *semistable* (this concept will be discussed below). For technical reasons, we will also assume that $p > 3$.

Although Frey had published a paper about Frey curves in 1982 [**8**], things didn't get really interesting until 1985, when Frey tried to prove that the Taniyama-Shimura Conjecture implies FLT. But his proof had some serious gaps. Several people tried to fix Frey's argument, and it was Jean-Pierre Serre [**24**] who saw that a special version of a conjecture he made on level reduction for modular Galois representations would fill the gap. Hence we may credit Frey and Serre with showing that FLT follows from Taniyama-Shimura and the special level reduction conjecture made by Serre. Versions of this argument can be found in [**7**] and [**22**] and one should also consult Serre's article [**25**].

Then, in 1986, Ken Ribet made significant progress along this route to FLT by proving this version of Serre's conjecture, and his proof eventually appeared in [**22**]. Thus FLT (for *all* primes p) was now a consequence of the Taniyama-Shimura Conjecture! Inspired by this development, Andrew Wiles began to work on Taniyama-Shimura, and seven years later, he presented a proof on June 23, 1993 that the conjecture is true for semistable elliptic curves, which (as we will see below) is good enough to prove FLT. Wiles' argument is not easy—the manuscript containing the proof is over 200 pages long. But many people in the mathematical community are confident that the proof will hold up under careful scrutiny. For a broad outline of Wiles' argument, see Ribet's article [**23**] (this article also has some useful references).

One interesting observation is that Frey was not the first to discover the Frey curve. On page 262 of [**12**], Hellegouarch writes down the Frey curve for a solution to FLT of exponent $2p^h$. Frey curves also appear implicitly as part of the correspondence between Fermat curves and modular curves considered by Kubert and Lang [**15**]. But Frey was clearly the first to suspect that the Frey curve couldn't exist because of the Taniyama-Shimura Conjecture.

To explain the Taniyama-Shimura Conjecture, we first need to define the concept of modular function.

Definition. *A function $f(z)$ on the upper half plane $\{z = x + iy\colon\ y > 0\}$ is* a modular function of level N *if $f(z)$ is meromorphic, even at the cusps (see the sidebar "The modular curve $X_0(N)$"), and for all integers a, b, c, d with $ad - bc = 1$ and $N|c$, we have*

$$f\left(\frac{az + b}{cz + d}\right) = f(z).$$

Taniyama-Shimura Conjecture. *Given an elliptic curve $y^2 = Ax^3 + Bx^2 + Cx + D$ over* **Q**, *there are nonconstant modular functions $f(z)$, $g(z)$ of the same level N such that*

$$f(z)^2 = Ag(z)^3 + Bg(z)^2 + Cg(z) + D.$$

Thus the Taniyama-Shimura Conjecture says that an elliptic curve over **Q** can be parameterized by modular functions, or, as Mazur says in [**19**], it has a "hyperbolic uniformization." Such an elliptic curve is said to be *modular*. Wiles

The modular curve $X_0(N)$

In the text, we considered the transformations $z \mapsto (az + b)/(cz + d)$ of the upper half plane $\mathfrak{h} = \{z = x + iy: y > 0\}$ associated to the group of matrices

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a, b, c, d \in \mathbf{Z}, ad - bc = 1, N|c \right\}.$$

This group acts on $\mathfrak{h}$ with quotient $\mathfrak{h}/\Gamma_0(N)$, and there is a compact Riemann surface $X_0(N)$ such that

$$\mathfrak{h}/\Gamma_0(N) = X_0(N) - \{\text{finite set of points}\}.$$

These points are the *cusps* and $X_0(N)$ is a *modular curve of level N*. (Other modular curves come from using different groups of matrices).

A function $f(z)$ is invariant under $\Gamma_0(N)$ iff it descends to a function on $\mathfrak{h}/\Gamma_0(N)$. Thus the definition of modular function means that we have a meromorphic function on $X_0(N)$. Furthermore, the Taniyama-Shimura Conjecture asserts that if E is an elliptic curve over $\mathbf{Q}$, then there is a surjective holomorphic map $X_0(N) \to E$.

As suggested in the text, a holomorphic 1-form on $X_0(N)$ is written $F(z)\,dz$, where $F(z)$ is a cusp form of weight 2 and level N. It follows that the genus of $X_0(N)$ (see the sidebar "The genus of an algebraic curve") equals the dimension of the space of these cusp forms.

It is known that $X_0(2)$ has genus zero (this follows by looking at the fundamental domain of $\Gamma_0(2)$ acting on $\mathfrak{h}$). Hence *there are no cusp forms of weight* 2 *and level* 2. This fact is used in the proof of FLT.

The modular curves $X_0(N)$ play an important role in the theory of elliptic curves. Some basic facts about modular curves can be found in **[14]** and **[26]** (and other references can be found in these books).

proved this conjecture for semistable elliptic curves. We should mention that our statement of the conjecture is very naive—some work is needed to show it is equivalent to the usual formulation (see the technical appendix to **[19]**). For a discussion of the conjecture and some of its history, see pages 130–135 of Lang's book **[17]**. At a more elementary level, Mazur's article "Number theory as gadfly" **[19]** gives a lovely introduction to the Taniyama-Shimura Conjecture.

Besides modular functions, we also need to know about modular forms of weight 2. The easiest way to see how these arise is through elliptic integrals. An elliptic integral is an integral of the form

$$\int \frac{dx}{\sqrt{Ax^3 + Bx^2 + Cx + D}}.$$

(Strictly speaking, this is only an elliptic integral of the first kind—there are many other types of elliptic integrals.) If $y^2 = Ax^3 + Bx^2 + Cx + D$, then this integral is simply $\int dx/y$. For a modular elliptic curve, we have $x = f(z)$, $y = g(z)$, and then

$$\frac{dx}{y} = \frac{df}{g} = \frac{f'(z)\,dz}{g(z)} = F(z)\,dz.$$

One can show that for a, b, c, d as in the definition of modular function, the above function $F(z)$ transforms via the rule

$$F\left(\frac{az+b}{cz+d}\right) = (cz+d)^2 F(z).$$

We call $F(z)$ *a modular form of weight* 2 *and level* N. When the modular parametrization is chosen correctly, the function $F(z)$ has some remarkable properties. It is holomorphic and vanishes at the cusps, and for this reason is called a *cusp form*. In addition, $F(z)$ is an *eigen-form* for the action of a certain Hecke algebra on the space of all cusp forms. So $F(z)$ is a rather sophisticated object.

The miracle is that $F(z)$ is intimately connected to the curve $y^2 = Ax^3 + Bx^2 + Cx + D$. Roughly speaking, one can reconstruct $F(z)$ simply by knowing the number of solutions of the congruences $y^2 \equiv Ax^3 + Bx^2 + Cx + D \bmod p$ for all primes p. Then the fact that $F(z)$ is a cusp form of weight 2 and level N tells us some profound things about the elliptic curve. This is one reason why Taniyama-Shimura is such a wonderful conjecture—number theorists would be excited by its proof even if there were no connection to FLT.

We can now sketch the argument of Frey and Serre which shows why FLT follows from Taniyama-Shimura and the level reduction conjecture of Serre. We begin with the FLT solution $a^p + b^p = c^p$. As above, we will assume $p > 3$ is prime and a, b, c are relatively prime with b even and $a \equiv -1 \bmod 4$. We then get the Frey curve $y^2 = x(x - a^p)(x + b^p)$.

The *discriminant* of a polynomial is the product of the squares of the differences of its roots. For the cubic $x(x - a^p)(x + b^p)$, the discriminant equals

$$(a^p - 0)^2(-b^p - 0)^2(a^p - (-b^p))^2 = a^{2p}b^{2p}c^{2p}$$

Invariants of the Frey curve

Suppose that we have the Frey curve $y^2 = x(x - a^p)(x + b^p)$ with our usual assumptions on a, b, c. Then we get the following invariants:

- Besides the discriminant defined in the text, an elliptic curve over $\mathbf{Q}$ has a more subtle invariant called the *minimal discriminant*. The minimal discriminant of the Frey curve is $\Delta = 2^{-8}a^{2p}b^{2p}c^{2p}$. Since b is even and $p \geq 5$, this is still an integer. This differs from the discriminant because the discriminant depends on the particular equation defining the curve, while the minimal discriminant is intrinsic to the curve.
- The *conductor* of the Frey curve is $N = \prod_{l|abc} l$. The conductor is the most subtle of the invariants associated to an elliptic curve over $\mathbf{Q}$. One can show that a modular elliptic curve is parametrized by modular functions whose level N equals the conductor of the curve.
- The *j-invariant* of the Frey curve is $j = 2^8(a^{2p} + b^{2p} + a^p b^p)^3/a^{2p}b^{2p}c^{2p}$. The j-invariant classifies the curve up to isomorphism over the complex numbers.

Precise definitions of these invariants can be found in Silverman's book [**26**] (see pp. 48, 224 and 361). The calculations for the Frey curve can be found in [**7**] and [**25**].

since a, b, c is a solution to FLT. It is unusual for a discriminant to be a pure $2p^{\text{th}}$ power—this is our first hint that the Frey curve is very special.

Besides the discriminant of its equation, an elliptic curve over $\mathbf{Q}$ has a variety of invariants, including its *minimal discriminant* Δ, *conductor* N and *j-invariant* j. For the Frey curve, these invariants are given in the sidebar "Invariants of the Frey curve." In general, Δ, N and j give useful information about the elliptic curve. For instance, when the curve is modular, one can find a modular parametrization using modular functions of level N, where N is the conductor of the curve. This fact will play an important role in the proof below.

We then have the following results about the Frey curve:

Lemma 1. *The Frey curve is semistable.*

Proof: We first need to define what semistable means. When a prime l divides the discriminant, two or possibly all three of the roots become congruent modulo l. Roughly speaking, an elliptic curve is semistable if for all such primes l, only two roots become congruent mod l (the definition is more complicated for the primes 2 and 3). Thus, for primes bigger than 3, the Frey curve is semistable since the discriminant is $a^{2p}b^{2p}c^{2p}$ and the roots are 0, a^p and $-b^p$, where a^p and b^p are relatively prime. More work is required to check semistability at $l = 2$ or 3, and when $l = 2$, the conditions b even, $a \equiv -1 \bmod 4$ and $p > 3$ are needed. For more details, see [7] and [25]. Q.E.D.

Corollary (Wiles). *The Frey curve is modular.*

Lemma 2. *For every odd prime l dividing N, the j-invariant of the Frey curve can be written as $j = l^{-mp} \cdot q$, where m is a positive integer and q is a fraction not involving l. (We say that the j-invariant is* exactly divisible *by l^{-mp} in this case.)*

Proof: The j-invariant of the Frey curve is

$$\frac{2^8(a^{2p} + b^{2p} + a^p b^p)^3}{a^{2p}b^{2p}c^{2p}} = \frac{2^8(c^{2p} - b^p c^p)^3}{(abc)^{2p}}.$$

The power of l dividing the denominator is obviously a multiple of p, and since a, b, c are relatively prime, one sees that $(c^{2p} - b^p c^p)^3$ and $(abc)^{2p}$ are relatively prime. Since N is the product of the primes dividing abc, the lemma follows easily. The lemma fails for $l = 2$ because of the factor of 2^8 in numerator. Q.E.D.

In the context of these three results—semistable modular elliptic curves whose j-invariants are exactly divisible by $l^{-\text{multiple of } p}$ for odd primes l dividing N—the level reduction conjecture of Serre applies for *all* odd primes dividing N (see [6] and [25] for the details of how this works). Serre's conjecture involves Galois representations and is rather technical (see [22] for a precise statement), though we will discuss its implications below.

We can now prove Fermat's Last Theorem:

Theorem. *The equation $x^p + y^p = z^p$ has no solutions with a, b, c nonzero for p an odd prime.*

Proof: Suppose there were a solution $a^p + b^p = c^p$, with our usual assumptions about p and a, b, c. Then we have a Frey curve, which by the above corollary has a cusp form F of weight 2 and level N, where N is the conductor. The Frey curve also has a Galois representation ρ on the points of order p on the curve (we won't define precisely what this means). The cusp form F is linked to the representation ρ in an especially nice way.

Serre's level reduction conjecture deals with the pair (ρ, F), and as we observed above, the hypotheses of the conjecture are satisfied for all odd primes l dividing N. In such a case, the conjecture asserts that there is a cusp form F' of weight 2 and level N/l with

$$F' \equiv F \bmod p$$

and F' is also an eigen-form for the appropriate Hecke algebra (it takes some work to define what it means for modular forms to be congruent modulo p). This congruence means that F' is linked to ρ in the same way F was, except that F' has smaller level N/l. But then, if l' is another odd prime dividing N, we can apply the level reduction conjecture to the pair (ρ, F') and get a cusp form F'' with even smaller level N/ll', and then apply it again to (ρ, F''), etc. Eventually we get a cusp form $\tilde{F}$ of weight 2 and level 2. (Note that 2 divides the conductor N since b is even.) Here is a diagram of the argument so far:

$$\begin{aligned} \text{solution of FLT} \rightarrow \text{Frey curve} &\rightarrow \text{cusp form of level } N \\ &\rightarrow \text{cusp form of level } N/l \\ &\rightarrow \text{cusp form of level } N/ll' \\ &\ \ \vdots \\ &\rightarrow \text{cusp form of level 2} \end{aligned}$$

But it is well known that there are no cusp forms of weight 2 and level 2 (see the sidebar "The modular curve $X_0(N)$"). Hence the above diagram self-destructs, and Fermat's Last Theorem is proved! Q.E.D.

This brings us to the end of the article, but certainly not to the end of the story. One thing missing from this account of Fermat's Last Theorem is the work of the many mathematicians who created the theories of elliptic curves, modular forms and Galois representations, and searched out the amazing connections between them. There is a lot more to say about the mathematics involved in the proof of Fermat's Last Theorem!

For an introduction to some of this wonderful material, we recommend the books by Koblitz **[13]**, Knapp **[14]** and Silverman **[26]**. At the undergraduate level, the recent book by Silverman and Tate **[27]** discusses elliptic curves and introduces the idea of a Galois representation.

NOTE ADDED IN PROOF: As of December 1993, Wiles's manuscript has not yet been released. Ken Ribet notes that a delay like this is relatively normal in connection with a long manuscript. Most experts continue to believe in the fundamental correctness of the proof.

REFERENCES

1. L. Adelman, D. Heath-Brown, The first case of Fermat's last theorem, *Invent. Math.* 79(1985), 409–416.
2. J. Buhler, R. Crandall, R. Ernvall and T. Metsänkylä, Irregular primes and cyclotomic invariants to four million, *Math. Comp.* 60 (1993), 151–153.
3. H. Edwards, *Fermat's Last Theorem*, Springer-Verlag, Berlin Heidelberg New York, 1977.
4. L. Euler, *Elements of Algebra*, Fifth Edition, Translated by J. Hewlett, Longman, Orme and Co., London, 1840. (Reprint by Springer-Verlag, Berlin Heidelberg New York, 1984.)

5. G. Faltings, Eindlichkeitssätze für abelschen Varietäten über Zahlköpern, *Invent. Math.* 73 (1983), 349–366. For an English translation, see *Arithmetic Geometry*, edited by G. Cornell and J. Silverman, Springer-Verlag, Berlin Heidelberg New York, 1986, 9–27.
6. de Fermat, Pierre, *Oeuvres de Fermat*, Volume 3, Gauthier-Villars, Paris, 1896.
7. G. Frey, Links between stable elliptic curves and certain diophantine equations, *Ann. Univ. Sarav.* 1 (1986), 1–40.
8. G. Frey, Rationale Punkte auf Fermatkurven und getwisteten Modulkurven, *J. reine u. angew. Math.* 331 (1982), 185–191.
9. A. Granville, The set of exponents, for which Fermat's Last Theorem is true, has density one, *Comptes Rendus/Mathematical Reports*, Academy of Science, Canada, 7 (1985), 55–60.
10. T. Heath, *Diophantus of Alexandria*, Second Edition, Cambridge University Press, Cambridge, 1910. (Reprint by Dover Books, New York, 1964.)
11. D. Heath-Brown, Fermat's Last Theorem is true for "almost all" exponents, *Bull. Lon. Math. Soc.* 17 (1985), 15–16.
12. Y. Helloguarch, Points d'ordre $2p^h$ sur les courbes elliptiques, *Acta Arith.* 26 (1975), 253–263.
13. N. Koblitz, *Introduction of Elliptic Curves and Modular Forms*, Springer-Verlag, Berlin Heidelberg New York, 1984.
14. A. Knapp, *Elliptic Curves*, Princeton University Press, Princeton, 1992.
15. D. Kubert and S. Lang, Units in the modular function field, I, *Math. Ann.* 218 (1975), 67–96.
16. S. Lang, *Introduction to Arakelov Theory*, Springer-Verlag, Berlin Heidelberg New York, 1988.
17. S. Lang, *Number Theory III: Diophantine Geometry*, Springer-Verlag, Berlin Heidelberg New York, 1991.
18. S. Lang, Old and new conjectured diophantine inequalities, *Bull. AMS* 23 (1990), 37–75.
19. B. Mazur, Number theory as gadfly, *Am. Math. Monthly* 98, 593–610.
20. A. Parshin, The Bogomolov-Miyaoka-Yau inequality for the arithmetical surfaces and its applications, *Séminaire de Théorie des Nombres*, Paris, 1986–87.
21. P. Ribenboim, 13 *Lectures on Fermat's Last Theorem*, Springer-Verlag, Berlin Heidelberg New York, 1979.
22. K. Ribet, On modular representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ arising from modular forms, *Invent. Math.* 100 (1990), 431–476.
23. K. Ribet, Wiles proves Taniyama's Conjecture; Fermat's Last Theorem follows, *Notices AMS* 40 (1993), 575–576.
24. J.-P. Serre, Lettre à J.-F. Mestre, in *Current Trends in Arithmetical Algebraic Geometry*, Contemporary Mathematics 67, AMS, Providence, 1987, 263–268.
25. J.-P. Serre, Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, *Duke Math. J.* 54 (1987), 179–230.
26. J. Silverman, *The Arithmetic of Elliptic Curves*, Springer-Verlag, Berlin Heidelberg New York, 1986.
27. J. Silverman and J. Tate, *Rational Points on Elliptic Curves*, Springer-Verlag, Berlin Heidelberg New York, 1992.
28. S. Wagon, The evidence: Fermat's Last Theorem, *Math. Intelligencer* 8, No. 1 (1986), 59–61.

Department of Mathematics & Computer Science
Amherst College
Amherst, MA 01002
dac@cs.amherst.edu

Who Was the Author?

Zur Invariantentheorie der Formen von n Variabeln, J. Ber. d. DMV, 1910.

Idealtheorie in Ringbereichen, Math. Ann., 1921.

Hilberstsche Anzahlen in der Idealtheorie, J. Ber. d. DMV, 1925.

Nichtkommutative Algeben, Math. Z. 1933.

Answer on page 54