MODULAR FORMS AND DIOPHANTINE QUESTIONS

KENNETH A. RIBET
Mathematics Department
University of California
Berkeley, CA 94720-3840
USA
e-mail: ribet@math.berkeley.edu

This article discusses many of the topics that I touched on during my Public Lecture at the National University of Singapore and my Lecture to Schools at Victoria Junior College. During the former lecture, I spoke in broad terms about the history of Fermat's Last Theorem and about the connection between Fermat's Last Theorem, and the conjecture — now a theorem! — to the effect that elliptic curves are related to modular forms. In my Lecture to Schools, I discussed questions that have been sent to me by students and amateur mathematicians.

1 Introduction

I have already written about Fermat's Last Theorem on a number of occasions. My article [25] with Brian Hayes in *American Scientist* focuses on the connection between Fermat's equation and elliptic curves. It was written in 1994, when the proof that Andrew Wiles announced in 1993 was not yet complete. My exposition [23] is intended for professional mathematicians who are not necessarily specialists in number theory. The introduction [27] by Simon Singh and me will be useful to readers who seek a summary of Singh's book [26] and to the documentary on Fermat's Last Theorem that Singh directed for the BBC [17].

I hope that the present article will offer a useful further look at some of the mathematics associated with Fermat's Last Theorem.

2 Background

Arguably the single most famous statement in mathematics is the assertion that Fermat's equation

$$a^n + b^n = c^n$$

has no solutions in positive integers a, b, and c when n is an integer greater than 2. According to his son Samuel, Pierre de Fermat wrote this assertion in the margin of his copy of Diophantus's *Arithmetic*, roughly in 1637.

Although Fermat may have believed in the 1630's that he had a proof of what came to be known as "Fermat's Last Theorem," we can only speculate as to what Fermat had in mind. It is widely believed that the argument that Fermat had mapped out for himself ran into unexpected difficulties. Indeed, when he was a mature mathematician, Fermat detailed a proof that

$$a^4 + b^4 = c^2$$

has no solution in positive integers, thus proving in particular that a perfect fourth power is not the sum of two others. Had Fermat been able to treat $a^n + b^n = c^n$ for all n, he probably would not have been interested in the special case $n = 4$.

It is worth pointing out that Fermat made at least one other mathematical assertion that proved to be incorrect: Fermat believed that the "Fermat numbers" $F_n := 2^{2^n} + 1$ are all prime. The first few of them — 3, 5, 17, 257 and 65537 — are indeed prime numbers. The next number in the series, $F_5 = 2^{32} + 1 = 4294967297$, is *not* a prime: it's the product of 641 and 6700417. Incidentally, there is no known n bigger than 4 for which F_n is prime. On the other hand, the numbers $F_6, \ldots, F_{24}$ are known currently to be composite (i.e., non-prime).

3 Early History

Fermat's Last Theorem has a long history, beginning with Fermat's work on the case $n = 4$ and Euler's 18th century study of $a^3 + b^3 = c^3$.

The techniques used in the 17th and 18th centuries are now included in the curriculum of undergraduate courses in number theory. For example, the work of Fermat and of Euler is discussed in the first two chapters of [11] and at various junctures in [14]. (The latter book is one of my favorite introductions to number theory. I recommend it enthusiastically to Berkeley students who seek an introduction to modern methods in number theory.)

After thinking about the first cases $n = 3$ and $n = 4$ of Fermat's equation, one turns naturally to exponents larger than 4. In fact, a simple remark shows that one need treat only the case where n is a prime number bigger than 2. Indeed, it is clear that Fermat's assertion, when true for a given exponent n, is true for all exponents that are multiples of n. For example, knowing the assertion for $n = 3$ allows us to conclude that here are no counterexamples to Fermat's assertion when n is 6, 9, 12, and so on. This remark follows from the simple observation that any perfect sixth power is in particular a perfect cube, and so forth.

Now any integer n bigger than 2 is either a power of 2 (2^t with $t \geq 2$) or else is a multiple of some prime number $p > 2$. Since integers of the first kind are divisible by 4 — an exponent for which Fermat himself proved Fermat's Last Theorem — it suffices to consider exponents that are odd prime numbers when one seeks to prove Fermat's Last Theorem. In other words, after verifying Fermat's assertion for $n = 4$ and $n = 3$, mathematicians were left with the problem of proving the assertion for the exponents 5, 7, 11, 13, 17, and so on.

Progress was slow at first. The case $n = 5$ was settled by Dirchlet and Legendre around 1825, while the case $n = 7$ was treated by Lamé in 1832.

In the middle of the nineteenth century, E. Kummer made a tremendous advance by proving Fermat's Last Theorem for an apparently large class of prime numbers, the *regular primes.* The definition of this class may be given quickly, thanks to a numerical criterion that was established by Kummer. Namely, one considers the expression

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \frac{x^2}{12} - \frac{x^4}{720} + \frac{x^6}{30240} - \frac{x^8}{1209600} + \frac{x^{10}}{47900160} - \frac{691\, x^{12}}{1307674368000} + \cdots$$

and defines the ith Bernoulli number B_i to be the coefficient of $\frac{x^i}{i!}$ in this expansion. Thus B_{12}, for example is $-\frac{691}{2730}$; the denominator is $2 \cdot 3 \cdot 5 \cdot 7 \cdot 13$, the product of those primes p for which $p-1$ divides 12. A prime number $p \geq 7$ is regular if p divides the numerator of none of the even-indexed Bernoulli numbers $B_2, B_4, \ldots, B_{p-3}$. The primes $p < 37$ turn out to be regular. On the other hand, 37 is irregular (i.e., not regular) because it divides the numerator of B_{32}: the numerator is $7709321041217 = 37 \cdot 683 \cdot 305065927$. (We may conclude that 683 and 305065927 are irregular as well.)

A proof of Fermat's Last Theorem for regular primes, along the lines given by Kummer, may be found in [11, Ch. 5]. See also [19] and [2] for alternative discussions. In these books, the reader will find a proof that there are infinitely many irregular prime numbers; see, for example, [19, Ch. VI, §4] or [2, Ch. 5, §7.2]. Although heuristic probabilistic arguments suggest strongly that regular primes should predominate, the set of regular primes is currently not known to be infinite.

Over the years, Kummer's work was refined repeatedly. Aided by machine calculation, mathematicians employed criteria such as those presented in [19] to verify Fermat's Last Theorem for all prime exponents that did not exceed ever increasing bounds. Most notably, four mathematicians proved Fermat's Last Theorem for all prime exponents below four million in an article that was published in 1993 [4]. It is striking that the calculations in that article were motivated by questions involving Bernoulli numbers and the arithmetic of cyclotomic fields; the proof of Fermat's Last Theorem for a large set of prime numbers came almost as an afterthought.

Readers used to dealing with experimental sciences might well now ask why a mathematician would insist on a rigorous proof of a statement, depending on a parameter n, that can be verified by calculation for all $n \leq 4,000,000$. A statement that is true in this range seems very likely to be true for all n. To answer this question, it suffices to point out the logical possibility that an assertion that is true experimentally may have one or more counterexamples that happen to be very large.

In fact, assertions that realize this possibility are not hard to find in number theory. As Fermat himself knew, the first solution to $x^2-109y^2 = 1$ in positive integers x and y is given by $x = 158070671986249$, $y = 15140424455100$. (See [30, Ch. II, §XII] for an illuminating discussion of Fermat's study of $x^2 - Ny^2 = \pm 1$.) If we set out to examine $x^2 - 109y^2 = 1$ with a computer, we might look for solutions with x and y non-zero, find no such solutions, and conclude incorrectly that this equation has only the trivial solutions $(-1, 0)$ and $(1, 0)$.

Here's another example: Euler conjectured in the eighteenth century that a perfect fourth power cannot be the sum of three perfect fourth powers. Noam Elkies [12] found the first counterexample to Euler's conjecture in 1988:

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

These examples illustrate the fact that numerical evidence in number theory can be misleading.

4 Modern History

The proof of Fermat's Last Theorem at the end of the last century hinges on a connection between putative solutions of Fermat's equation and cubic equations with integer coefficients (elliptic curves). To have a solution to Fermat's equation is to have positive integers a and b for which $a^n + b^n$ is a perfect nth power. (We shall suppose that n is at least 5 and that n is a prime number. The results of Fermat and Euler imply that these assumptions are harmless.) Given a and b, we consider the equation

$$E : y^2 = x(x - a^n)(x + b^n),$$

in which x and y are new variables. This equation defines an elliptic curve.

The connection between Fermat and elliptic curves was noticed by several mathematicians, including Yves Hellegouarch and Gerhard Frey. In a recent book [13], Hellegouarch recounts the history of this connection. It was Frey who had the decisive idea that E could not possibly satisfy the *Shimura–Taniyama conjecture*, which states that elliptic curves are *modular*. (We shall discuss this crucial property in §5 below.)

Frey's suggestion became known to the mathematical community in the mid 1980s. In 1986, I proved that elliptic curves associated solutions to Fermat's equation are non-modular, thereby showing that Fermat's Last Theorem is a consequence of the Shimura–Taniyama conjecture [21], [22]. Said differently: each solution to Fermat's Last Theorem gives a counterexample to the Shimura–Taniyama conjecture. Thus if that conjecture is true, so is Fermat's Last Theorem.

As the reader is no doubt aware, Andrew Wiles worked in his Princeton attic from 1986 to 1993 with the goal of establishing the Shimura–Taniyama conjecture. Although the conjecture per se was a central problem of number theory, Wiles has stated that he was drawn to this problem because of the link with Fermat's Last Theorem. In June, 1993, Wiles announced that he could prove the Shimura–Taniyama conjecture for a wide class of elliptic curves, including those coming from Fermat solutions. This announcement implied that the proof of Fermat's Last Theorem was complete.

After a short period of celebration among mathematicians, Wiles's colleague Nicholas Katz at Princeton found a "gap" in Wiles's proof. Because the gap's severity was not appreciated at first, it was months before the existence of the gap was known widely in the mathematical community. By the end of 1993, however, the fact that Wiles's proof was incomplete was reported in the popular press.

The proof announced by Wiles remained in doubt until October, 1994, when Richard Taylor and Andrew Wiles released a modified version of the proof that circumvented the gap. The new proof was divided into two articles, one by Wiles alone and one a collaboration by Taylor and Wiles [32], [29]. The two articles were published together in 1995. The proof presented in those articles was accepted quickly by the mathematical community.

As a result of his work, Wiles has been honored repeatedly. For example, in December, 1999, he was knighted by the Queen: he received the "KBE/DBE" along with Julie Andrews, Elizabeth Taylor and Duncan Robin Carmichael Christopher, Her Majesty's ambassador to Jakarta[a].

After the manuscripts by Wiles and Taylor–Wiles were written in 1994, the technology for establishing modularity became increasingly more sophisticated and more general. The class of curves to which the technology can be

[a] `http://files.fco.gov.uk/hons/honsdec99.shtml`

applied was enlarged in three stages [10], [5], [3]. In the last stage, four mathematicians — Christophe Breuil, Brian Conrad, Fred Diamond and Richard Taylor — announced in June, 1999 that they had proved the full Shimura–Taniyama conjecture, i.e., the modularity of all elliptic curves (that are defined by equations with integer coefficients). Although their proof is not yet published, it is available from `http://www.math.harvard.edu/~rtaylor/`, Richard Taylor's Web site at Harvard. In addition, the proof has been the subject of a substantial number of oral presentations. In particular, the proof was explained by the four authors in a series of lectures at a conference held at the Mathematical Sciences Research Institute in Berkeley, California in December, 1999.

5 The Shimura–Taniyama Conjecture

The conjecture hinges on the notion of "arithmetic mod p," p being a prime number. When working mod p, we ignore all integers that are multiples of p. In other words, when we interact with an integer m, we care only about the remainder when m is divided by p. This remainder is one of the numbers 0, 1, 2,. . . , $p - 1$. For example, the integers mod 5 are 0, 1, 2, 3 and 4.

Suppose that we are given an equation with integer coefficients. Then for each prime number p, we can use the equation to define a relation mod p. As an illustration, the simple equation $x^2 + y^2 = 1$ gives rise to a relation mod 2, mod 3, mod 5, and so on.

This type of relation is best illustrated by a concrete example. Suppose that we take $p = 5$, so that the numbers mod 5 are the five numbers that we listed above. There are thus 25 pairs of numbers (x, y) mod 5. For each pair (x, y), we can ask whether $x^2 + y^2$ is the same as 1 mod 5. For $(0, 4)$, the answer is "yes" because 16 and 1 are the same mod 5. For $(2, 2)$, the answer is "no" because 8 and 1 are not the same mod 5. After some calculation, one finds that there are four pairs of numbers mod 5 for which the answer is in the affirmative. These pairs are $(0, 1)$, $(0, 4)$, $(1, 0)$ and $(4, 0)$. After we recognize that 4 is that same as -1 mod 5, we might notice that the four solutions that we have listed have analogues for every prime number $p > 2$. There are always the four systematic solutions $(0, 1)$, $(0, -1)$, $(1, 0)$ and $(-1, 0)$ for each such prime.

We can make a similar calculation mod 7. It is fruitful to begin by listing the squares of the seven numbers mod 7:

a	0	1	2	3	4	5	6
a^2	0	1	4	2	2	4	1.

In which ways can we write 1 as the sum of two numbers in the bottom row (possibly the sum of a number and itself)? We can write 1 as $0+1=1+0$, and we can also write 1 as $4+4$ (since 8 is the same as 1 mod 7). We end up with the four "new" solutions $(\pm 2, \pm 2)$ in addition to the four systematic solutions that we listed in connection with the case $p=5$. As a consequence, there are eight solutions to $x^2+y^2=1$ mod 7.

After experimenting with other primes ($p=11$, $p=13$, etc.), you will have little trouble guessing the general formula for the number of solutions to $x^2+y^2=1$ mod p. When $p=2$, there are the two solutions $(0,1)$ and $(1,0)$. When p is bigger than 2, there are either $p+1$ or $p-1$ solutions to $x^2+y^2=1$ mod p, depending on whether p is 1 less than or 1 more than a multiple of 4. This simple recipe was known centuries ago. It can be established in various ways; perhaps I should leave its proof as an exercise for the interested reader.

The equation $x^2+y^2=1$ was intended as a warm-up; we shall now consider the superficially analogous equation $x^3+y^3=1$. Here again we study the number of solutions to $x^3+y^3=1$ mod p and seek to understand how this number varies with p. It turns out that the quantity p mod 3 plays an important role here — just as the behavior of p mod 4 was significant for $x^2+y^2=1$. When $p=3$, the quantity x^3 mod p coincides with x mod 3; this is a special case of what is called "Fermat's Little Theorem" in textbooks. Hence the solutions to $x^3+y^3=1$ mod 3 are the same as the solutions to $x+y=1$; there are three solutions, because x can be taken arbitrarily, and then y is $1-x$ mod 3. If p is 2 mod 3, i.e., if p is 1 less than a multiple of 3, one shows by an elementary argument that there are again p solutions. (If p is 2 mod 3, then every number mod p has a unique cube root.)

The interesting case for this equation is the remaining case where p is 1 more than a multiple of 3. This case was resolved by Gauss in the nineteenth century. To see what is going on, we should look at a few examples:

First off, we take $p=7$. The cubes mod 7 are 0, 1 and $6=-1$. If two cubes sum to 0, one is 1 and the other is 0. Also, 1 has three cube roots: 1, 2 and 4. Thus there are six solutions to $x^3+y^3\equiv 1$, namely $(0,1)$, $(0,2)$, $(0,4)$ and the analogous pairs with x and y reversed.

When $p=13$, the cubes are 0, 1, -1, 5 and 8. There are again only six solutions because the only way to write 1 as a sum of two cubes is to take $0+1$ as before.

Now try $p=19$. It turns out that there are 24 solutions here—the six that we knew about already, together with 18 unexpected ones arising from the equation $1=8+(-7)$ and the fact that 8 and -7 are both cubes mod 19. (Since $4^3=64\equiv 7$ mod 19, -7 is the cube of -4.) We get 18 solutions by taking x to be one of the 3 cube roots of 8 and y to be one of the 3 cube roots of -7, or vice versa.

When $p = 31$, there are 33 solutions. (Note that $33 = 6 + 18 + 9$.) There are 6 solutions coming from $0 + 1 = 1$, 18 coming from $2 + (-1) = 1$ and 9 from $16 + 16 = 1$. Summary:

p	7	13	19	31	$\cdots$
# solns.	6	6	24	33	$\cdots$.

How does this table continue? What is the number of solutions that we get when p is, say, 103? It is hard to imagine the rule that expresses the number of solutions in terms of p.

Gauss found an expression for the number of solutions that we can view as a "generalized formula" [14, p. 97]. Namely, when $p \equiv 1 \bmod 3$, Gauss showed that one has

$$4p = A^2 + 27B^2$$

for some integers A and B. These integers are uniquely determined except for their signs. We can and do choose A so that $A \equiv 1 \bmod 3$. Then Gauss's formula states:

$$\#\ \text{solns.} = p - 2 + A.$$

For example, if $p = 13$, then $4p = 52 = 5^2 + 27 \cdot 1^2$. Thus $A = -5$. We have $p - 2 + A = 6$.

When $p = 31$, $4p = 124 = 4^2 + 27 \cdot 2^2$. Thus $A = 4$ and $p - 2 + A = 33$.

When $p = 103$, $4p = 13^2 + 27 \cdot 3^2$, so $A = 13$ and the number of solutions is 114.

The equation $x^3 + y^3 = 1$ defines one of the simplest possible elliptic curves. Gauss's explicit recipe shows in particular that $x^3 + y^3 = 1$ defines a *modular* elliptic curve.

The Shimura–Taniyama conjecture states that there's an analogous "formula" for *every* elliptic curve. Because this formula involves modular forms, the Shimura–Taniyama conjecture is usually paraphrased as the statement that elliptic curves are modular.

For a random elliptic curve, the formula provided by the associated modular form is not as explicit as Gauss's formula for $x^3 + y^3 = 1$. Here is a famous example that *begins* to give the flavor of the general case: We consider first the formal power series with integral coefficients $\sum a_n X^n$ that is obtained by expanding out the product

$$X \prod_{m=1}^{\infty} (1 - X^m)^2 (1 - X^{11m})^2.$$

For all $n \geq 1$, a_n is an integer. In fact, the numbers a_n are the coefficients of the Fourier expansion of a well known modular form.

At the same time, we consider the elliptic curve defined by the equation $y^2 + y = x^3 - x^2$. Then a theorem of M. Eichler and G. Shimura states that, for each prime p (different from 11), the number of solutions to this equation mod p is $p - a_p$. The connection between the number of solutions and the pth coefficient of a modular form shows that the elliptic curve defined by $y^2 + y = x^3 - x^2$ is a modular elliptic curve. (A coffee mug that celebrates this relation is currently available from the Mathematical Sciences Research Institute. Go to `http://www.msri.org/search.html` and search for "coffee cup.")

6 Another Formula of Gauss

For a third example, we look at the elliptic curve defined by the equation $y^2 = x^3 - x$. Although its equation recalls the equation $y^2 + y = x^3 - x^2$ of the second example, this third example is much more analogous to the first example. To explain the analogy, it is important to recall a theorem of Fermat about sums of squares. Namely, suppose that p is a prime number and that we seek to write p is the form $r^2 + s^2$, where r and s are integers. If p is 2, we can write $p = 1^2 + 1^2$. If p is congruent to 3 mod 4, then it is *impossible* to write p as $r^2 + s^2$. Indeed, squares are congruent to either 0 or 1 mod 4; it is therefore impossible that a sum of two squares be congruent to 3 mod 4.

The interesting case is that where p is congruent to 1 mod 4, i.e., where p is 1 plus a multiple of 4. Fermat proved in that case that p may be written as a sum of two squares: we have $p = r^2 + s^2$ with r and s whole numbers. The pair (r, s) is clearly not unique because we can exchange r and s and we can change the signs of either or both of these integers. However, there is no more ambiguity than that: the integers r and s become unique up to sign after we require that r be odd and that s be even. Accordingly, r and s are determined completely if we require that r be odd, that s be even and that both integers be positive.

This theorem of Fermat is proved in most elementary number theory books; see, e.g., [14, Ch. 8] for one proof. (The uniqueness is left as an exercise at the end of the chapter.) A beautiful proof of the existence of r and s, due to D. Zagier, is presented in "Proofs from the Book" [1], a volume celebrating Paul Erdös's idea that there is frequently an optimally beautiful proof of a given proposition in mathematics.

Following Gauss, we will now adjust the sign of r (if necessary) to ensure that the sum $r + s$ is congruent to 1 mod 4. For example, suppose that $p = 5$, so that $(r, s) = (1, 2)$ under the initial choice that has both r and s positive. With this choice, $r + s = 3$ is *not* 1 mod 4. Accordingly, we change the sign of r and put $r = -1$. The sum $r + s$ is then 1, which of course is 1 mod 4. For another example, we take $p = 13$, so that $(r, s) = (3, 2)$ with the initial choice. Here $r + s = 3 + 2 = 5$, which is already 1 mod 4. We therefore leave r positive in this case. It is perhaps enlightening to tabulate the values of r and s for the first primes that are 1 mod 4. In doing so, we write "r_p" instead of "r" and "s_p" instead of "s" to stress that r and s depend on p:

p	5	13	17	29	37	41	53	61	73	89	$\cdots$
r_p	-1	3	1	-5	-1	5	7	-5	-3	5	$\cdots$
s_p	2	2	4	2	6	4	2	6	8	8	$\cdots$.

We return now to $y^2 = x^3 - x$ with the idea of calculating the number of solutions to the mod p congruence defined by this equation. If $p = 2$, there are two solutions: $(0, 0)$ and $(1, 0)$. If p is congruent to 3 mod 4, it turns out that there are exactly p solutions. More precisely, if x is 0, 1 or -1 (i.e., $p-1$) mod p, then $y = 0$ is the one value of y for which (x, y) is a solution. For each value of x different from 0 and ± 1, there are either two values of y or no values of y for which (x, y) is a solution mod p. (If a non-zero number mod p has a square root, it has exactly two square roots, which are negatives of each other.) An elementary argument shows that if there are two y for a given x, then there are no y for $-x$, and vice versa. The point here is that a non-zero number mod p has a square root mod p if and only if its negative does not; this observation is valid when p is 3 mod 4 but fails to be true when p is 1 mod 4. The end result is that, on average, there is one value of y that works for each x. Thus the number of solutions is p, as was stated.

The interesting case for $y^2 = x^3 - x$ is that where p is congruent to 1 mod 4. We assume now that this is the case. To get a feel for the situation, we can calculate the number of solutions mod 5 and mod 13; these are the first two primes that are 1 mod 4.

Suppose that $p = 5$. The values $x = 0$, $x = 1$ and $x = 4$ make $x^3 - x$ congruent to 0, so that they give rise to exactly one solution each; y must be 0. If $x = 2$, then $x^3 - x$ is congruent to 1, a number that has two square roots mod 5, namely ± 1. Thus $x = 2$ gives rise to two solutions. Similarly, if $x = 3$, then $x^3 - x$ is congruent to 4 mod 5, and 4 has two square roots. Thus $x = 2$ also gives rise to two solutions. As a result, there are seven solutions to $y^2 = x^3 - x$ mod 5.

Suppose now that $p = 13$. The three values $x = 0, 1, -1$ give rise to a single solution each as before; in each case, y is again 0. The ten remaining values of x (namely, $x = 2, 3, \ldots, 11$) each give rise either to two or to no solutions: the quantity $x^3 - x$ is non-zero mod p and we have to decide in each case whether or not it is a square (i.e., a number with square roots mod p). The quantities are respectively 6, 11, 8, 3, 2, 11, 10, 5, 2 and 7 mod 13. On the other hand, the non-zero squares mod 13 are 1, 3, 4, 9, 10 and 12. It happens, then, that only two of the numbers x between 2 and 11 are such that $x^3 - x$ is a square. Thus we find — one again — that there are seven solutions to $y^2 = x^3 - x$ mod p.

One could easily guess from these two examples that there are always seven solutions to $y^2 = x^3 - x$ mod p when p is 1 mod 4, but these two examples are misleading.

Theorem 1 (Gauss) *Suppose that p is a prime that is 1 mod 4. Then the number of solutions to $y^2 = x^3 - x$ mod p is $p - 2r_p$, where r_p is chosen as above.*

The theorem is compatible with the two examples that we presented. When $p = 5$, we have $r_p = -1$, so that $p - 2r_p = 7$. When $p = 13$, r_p is 3, and $13 - 2 \cdot 3 = 7$. Since $r_{73} = -3$, the number of solutions to $y^2 = x^3 - x$ mod 73 is $73 + 6 = 79$. Here is an example with a p that is considerably larger than the primes that have appeared thus far: Suppose that p is the prime number 144169. We can write p as the sum $315^2 + 212^2$. It follows that $r_p = \pm 315$. Since $315 + 212 = 527$ is 3 mod 4, we must take $r_p = -315$. Gauss's formula then asserts that the number of solutions to $y^2 = x^3 - x$ mod p is $144169 + 2 \cdot 315 = 144799$.

A variant of Gauss's formula is proved in [14, Ch. 11, §8]. The connection between the variant given there and the formula of Theorem 1 is made in Exercise 13 at the end of [14, Ch. 11].

7 Binomial Coefficients

Because I have written extensively about Fermat's Last Theorem, I have received a number of letters about number theory from amateur mathematicians. Several years ago, I received a letter about binomial coefficients. These are the numbers that appear in the expansion of $(x + y)^n$ when n is a positive integer. Recall, for example, that

$$\begin{aligned}
(x+y)^2 &= x^2 + 2xy + y^2, \\
(x+y)^3 &= x^3 + 3x^2y + 3xy^2 + y^3, \\
(x+y)^4 &= x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4, \\
(x+y)^5 &= x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5,
\end{aligned}$$

. . . .

The coefficient of $x^{n-i}y^i$ in the expansion of $(x+y)^n$ is usually denoted $\binom{n}{i}$. It may be expressed as the fraction $\frac{n!}{(n-i)!i!}$, where $k!$ is used to denote the product of the first k positive integers. (By convention, $0! = 1$.) Looking at the expansion for $(x+y)^4$, say, we see that $\binom{4}{2} = 6$ and that $\binom{4}{3} = 4$.

The letter that I received concerned the central coefficient in the expansion of $(x+y)^n$ when $n = (p-1)/2$ and p is a prime congruent to 1 mod 4. If p is 5, for example, this coefficient is 2. In general, it is the binomial coefficient $\binom{(p-1)/2}{(p-1)/4}$, a number that we can call b_p for short:

p	5	13	17	29	37	$\cdots$
b_p	2	20	70	3432	48620	$\cdots$.

These numbers grow large very quickly, but my correspondent was considering them modulo p in order to keep their size manageable. For reasons that I no longer recall, he hit upon the scheme of representing them modulo p as even numbers between $-(p-1)$ and $+(p-1)$. In the table that follows, I've written c_p for the unique even number in this range that is congruent to b_p modulo p:

p	5	13	17	29	37	41	53	$\cdots$
c_p	2	-6	2	10	2	10	-14	$\cdots$.

He noticed that the residues c_p were related to the integers r_p that we introduced above in connection with Gauss's formula. Because he did not have Gauss's formula in mind, he tabulated the r_p as *positive* numbers:

p	5	13	17	29	37	41	53	61	73	89	$\cdots$
c_p	2	-6	2	10	2	10	-14	10	-6	10	$\cdots$
r_p	1	3	1	5	1	5	7	5	3	5	$\cdots$.

It was clear to him empirically from his calculations that $c_p = \pm 2r_p$, but the sign in this equation seemed completely opaque. He asked me to determine the sign and to explain to him why the identity is true.

Although the identity $c_p = \pm 2r_p$ has been known at least since the nineteenth century, it was new to me. However, I realized that the sign that appears in this identity becomes significantly less mysterious once we again endow r_p with the sign that we introduced in connection with Gauss's formula:

p	5	13	17	29	37	41	53	61	73	89	$\cdots$
c_p	2	-6	2	10	2	10	-14	10	-6	10	$\cdots$
r_p	-1	3	1	-5	-1	5	7	-5	-3	5	$\cdots$.

The rule relating c_p and r_p in this latter table is as follows:

$$\begin{cases} c_p = +2r_p & \text{if } p \text{ is 1 plus a multiple of 8} \\ c_p = -2r_p & \text{if } p \text{ is 5 plus a multiple of 8.} \end{cases}$$

Notice here that p is assumed going in to be of the form $1 + 4t$. If t is even, p is congruent to 1 mod8; if t is odd, p is congruent to 5 mod 8.

It is not too hard to establish this rule if one takes Gauss's formula as a starting point. Indeed, suppose that we seek to calculate the number of mod p solutions to $y^2 = x^3 - x$. We can let x run over the set $\{0, 1, 2, \ldots, p-1\}$ of numbers mod p. For each x, the number of y satisfying $y^2 = x^3 - x$ is:

$$\begin{cases} 1 & \text{if } x^3 - x \text{ is 0 mod } p \\ 2 & \text{if } x^3 - x \text{ is a non-zero square mod } p \\ 0 & \text{if } x^3 - x \text{ is not a square mod } p. \end{cases}$$

This number may be written $1 + \left(\frac{x^3-x}{p}\right)$, where $\left(\frac{-}{p}\right)$ is the traditional Legendre symbol whose values are 0, $+1$, -1 according as the argument is 0, a non-zero square, or a non-square mod p. The number of solutions to $y^2 = x^3 - x$ mod p is then

$$\sum_{x=0}^{p-1} \left(1 + \left(\frac{x^3 - x}{p}\right)\right).$$

Thus

$$2r_p = -\sum_{x=0}^{p-1} \left(\frac{x^3 - x}{p}\right)$$

by Theorem 1. A standard congruence for $\left(\frac{-}{p}\right)$ states that $\left(\frac{a}{p}\right)$ is congruent mod p to $a^{(p-1)/2}$ for each integer a. Using this congruence, we get

$$2r_p \equiv -\sum_{x=0}^{p-1} (x^3 - x)^{(p-1)/2},$$

where "$\equiv$" denotes congruence mod p. The expression $(x^3 - x)^{(p-1)/2}$ can be expanded out as a sum that involves the binomial coefficients $\binom{(p-1)/2}{i}$

$(i = 0, \ldots, (p-1)/2)$. After changing the order of summation, we get from this expansion

$$2r_p \equiv -\sum_{i=0}^{(p-1)/2} (-1)^i \left\{ \sum_{x=0}^{p-1} \binom{(p-1)/2}{i} x^{3((p-1)/2)-2i} \right\}.$$

Now an elementary fact about sums of powers states that we have

$$\sum_{x=0}^{p-1} x^j \equiv \begin{cases} 0 & \text{if } p-1 \text{ does not divide } j \\ -1 & \text{if } p-1 \text{ does divide } j. \end{cases}$$

It follows that only one of the inner sums is non-zero mod p: this is the sum corresponding to the choice $i = (p-1)/4$. We thus get

$$2r_p \equiv -(-1)^{(p-1)/4} \binom{(p-1)/2}{(p-1)/4} (-1),$$

so that

$$2r_p \equiv (-1)^{(p-1)/4} b_p.$$

Since the sign in this expression is 1 if and only if p is congruent to 1 mod 8, we find that c_p is either $2r_p$ or $-2r_p$, with the choice of sign as stated above.

8 Sums of Squares mod p

During my "lecture to schools" at Victoria Junior College, I discussed several issues that were brought up by undergraduate students and amateur mathematicians. One was the binomial coefficient identity that is treated in the previous section — it turned out to be a corollary of a formula of Gauss that a number of mathematicians had been featuring in their lectures on Fermat's Last Theorem.

A second question involves squares mod p; it was posed by a freshman (i.e., first-year student) at Yale University. Recall from our discussion before that the non-zero squares mod 13 are 1, 3, 4, 9, 10 and 12. The sum of these six numbers, considered as positive integers, is $1 + 3 + 4 + 9 + 10 + 12 = 39$, which is $13 \cdot 3$. Suppose, more generally, that p is a prime different from 2. It is a standard fact from elementary number theory that there are precisely $(p-1)/2$ different non-zero squares mod p. (It's easy to see that there are *at most* this number because $(-x)^2$ is the same as x^2 for each x mod p. The point is that if $x^2 = y^2$, then $x \equiv \pm y$ mod p.) Regarding the squares as integers between 1 and $p-1$, we form their sum and call the resulting positive integer $S(p)$. Then $S(3) = 1$, $S(5) = 1 + 4 = 5$, $S(7) = 1 + 2 + 4 = 7$,

$S(11) = 1 + 3 + 4 + 5 + 9 = 22$, etc. We can guess from these examples that $S(p)$ is divisible by p for $p \geq 5$, and indeed this divisibility is relatively easy to establish. Let's assume then that p is at least 5 and set $L(p) = S(p)/p$. Thus $L(5) = L(7) = 1$, $L(11) = 2$, and so on.

The question concerns $L(p)$: can we find a formula for it?

p	5	7	11	13	17	19	23	29	31	37	41	43	47	53	59	61	67	71	$\cdots$
$L(p)$	1	1	2	3	4	4	4	7	6	9	10	10	9	13	13	15	16	14	$\cdots$

After looking at this table (or perhaps an extension of it), one comes to the realization that

$$L(p) = \frac{p-1}{4}$$

when p is congruent to 1 mod 4. This formula is given frequently as an exercise in number theory classes, and the proof is not difficult. The key fact in this case is that if a is a square mod p, then so is $-a$. (The analogous statement for p congruent to 3 mod 4 is that $-a$ is *never* a square mod p if a is a non-zero square!) Eqivalently, if a is a square between 1 and $p-1$, then $p-a$ is again such a number. Notice also that a and $p-a$ are distinct numbers since p is odd. Thus the squares mod p between 1 and $p-1$ can be partitioned into pairs $\{a, p-a\}$. The sum of the numbers in each pair is p. Since there are $(p-1)/2$ squares mod p, there are $(p-1)/4$ pairs. It follows that the sum of all the squares is $p\dfrac{p-1}{4}$, as was claimed. For example, if $p = 17$, then the squares are 1, 2, 4, 8, 9, 13, 15 and 16. We re-write the sum of these eight numbers as $(1+16)+(2+15)+(4+13)+(8+9) = 4 \cdot 17$.

Knowing the formula for $L(p)$ when $p \equiv 1 \bmod 4$, one might anticipate a similar formula in the complementary case $p \equiv 3 \bmod 4$. The values of $L(p)$ for $p = 7$, 11 and 19 suggest that one has

$$L(p) \overset{?}{=} \frac{p-3}{4}.$$

I was asked whether this formula was true in general.

In fact, I saw quickly that the formula is false by continuing the computation. Indeed, we have $L(23) = 4$ when the formula predicts $L(23) \overset{?}{=} 5$. On the other hand, the formula seemed perhaps to be not so far from the truth: in the table, the formula is correct for $p = 43$ and $p = 67$, as well as for the small values 7, 11 and 19. This behavior is somewhat striking, since false formulas tend to fail in a much more spectacular way.

Although I was initially puzzled by what was going on, I realized after orienting myself that I already knew how to write down a correct version of the formula. In fact, the information needed is available in a classic text on algebraic number theory, Hermann Weyl's "Algebraic Theory of Numbers" [31]. The corrected formula reads

$$L(p) = \frac{p-1-2h(p)}{4},$$

where $h(p)$ is a certain odd positive integer that depends on p. The fraction $(p-3)/4$ coincides with $(p-1-2h(p))/4$ if and only $h(p) = 1$.

The term $h(p)$ is a *class number* that has been studied extensively at least since the time of Gauss. One way to introduce the class number is to return to the proof of Fermat's Last Theorem for exponents 3 and 4. In the seventeenth century, Fermat proved that the equation $a^4 + b^4 = c^4$ has no solution in positive integers by proving a stronger statement involving squares and fourth powers. As recapitulated by Ireland and Rosen [14, p. 272], the idea is to show that there are no solutions to $a^4 + b^4 = c^2$ by a method of *descent*: one assumes that there is a solution to this equation in positive integers, chooses a ("minimal") solution for which the number c is as small as possible, and then parlays the chosen solution into a new solution whose c-value is even smaller than that for the minimal solution. The method hinges on properties of unique factorization of positive integers; unique factorization is the statement that an integer bigger than 1 can be written as a product of prime numbers in essentially only one way.

To exploit unique factorizaton, one re-writes the equation $a^4 + b^4 = c^2$ as the statement that $c^2 - a^4 = (c - a^2)(c + a^2)$ is a perfect fourth power (namely, b^4). One then recalls the principle that if a product of two positive integers $A \cdot B$ is a fourth power, then A and B must each be perfect fourth powers — provided that A and B have no common factor. The $(c - a^2)$ and $(c + a^2)$ may share a common factor, so that the principle cannot be used without modification. However, one can exchange a and b in the minimal solution (if necessary) so as to make 2 the only divisor > 1 that is common to the two factors $(c - a^2)$ and $(c + a^2)$. After using unique factorization in a judicious way, one emerges with a new solution whose c-value is smaller than the c-value of the minimal solution.

Euler treated the equation $a^3 + b^3 = c^3$ by a method that is similar to Fermat's. However, he needed to work with quantities that involve a complex cube root of 1 [14, Ch. 17, §8]. The key idea is to introduce quantities involving $\sqrt{-3}$ that act as generalized integers and to establish for these quantities an analogue of the unique factorization theorem for positive integers. These quantities are expressions of the form

$$n + m\frac{-1+\sqrt{-3}}{2},$$

where n and m are usual integers. These expressions can be added, subtracted and multipled to form an arithmetical system like the system of ordinary integers.

Suppose now more generally that p is a prime number that is congruent to 3 mod 4. Then one considers in an analogous way the system of quantities

$$n + m\frac{-1+\sqrt{-p}}{2},$$

where n and m are again integers. The class number $h(p)$ measures the extent to which unique factorization fails for this system; $h(p) = 1$ if and only if unique factorization can be established. Gauss proved that $h(p)$ is always an odd number and examined the behavior of $h(p)$ numerically. He conjectured that $h(p) = 1$ if and only if p is one of the prime numbers 3, 7, 11, 19, 43, 67, 163. This conjecture was established only in the twentieth century! See [28] for a discussion of work by A. Baker, K. Heegner and H. Stark that resolved this question. Thus the proposed formula is correct precisely for the values of p that we listed above — namely, 7, 11, 19, 43 and 67 — plus the larger prime $p = 163$.

9 Fermat-like Equations

At Victoria Junior College, I discussed a third question that was sent to me by an amateur mathematician: he asked whether the methods that proved Fermat's Last Theorem would shed any light on the Fermat-like equation $a^n + b^n = 2c^n$, where n is a positive integer. In this equation, we can suppose first that a, b and c are positive integers. We observe immediately that this equation, in contrast to Fermat's equation, does have solutions. Indeed, we can take a to be an arbitrary positive integer and set b and c equal to a! We can call these solutions trivial and ask whether there are non-trivial solutions to the equation. If $n = 1$, then the equation states simply that $c = a + b$, so of course there is no obstacle to having solutions. As with Fermat's equation, there are non-trivial solutions if $n = 2$; for example, we can take $(a, b, c) = (1, 7, 5)$. It is an interesting exercise to find a description of the non-trivial solutions that is analogous to the familiar description of all Pythagorean triples — the solutions to $a^2 + b^2 = c^2$.

The special appeal of $a^n + b^n = 2c^n$ stems from its reformulation as the statement that c^n is the average of a^n and b^n. In other words, if $a^n + b^n = 2c^n$, then a^n, c^n and b^n form an arithmetic progression. In the example $(1, 7, 5)$ with $n = 2$, 25 is the average of 1 and 49; equivalently, the differences $25 - 1$ and $49 - 25$ are equal.

As I gradually learned while researching this equation, its history is parallel to that of Fermat's equation. As with Fermat's equation, the discussion for exponents $n > 2$ may be reduced to the two cases $n = 4$ and $n = p$, where p is a prime greater than 2. Fermat proved that four distinct perfect squares cannot form an arithmetic progression, and he showed that there are no non-trivial solutions to $a^4 + b^4 = 2c^4$. Accordingly, it sufficed to consider the case $n = p$ with $p \geq 3$. Euler and Lengendre treated the case $p = 3$. In 1952, P. Dénes showed that there are no non-trivial solutions for $p \leq 29$ and conjectured that there are no non-trivial solutions for all prime exponents bigger than 2.

In an article that was published in 1997 [24], I adapted the technology that was used in proving Fermat's Last Theorem to establish Dénes's conjecture for primes p that are congruent to 1 mod 4. Subequently, H. Darmon and L. Merel settled Dénes's conjecture completely (in the affirmative) by introducing new techniques to deal with the case $p \equiv 3 \bmod 4$ [9]. There has been a substantial literature about Fermat-like equations ever since the connection between Fermat solutions and elliptic curves was uncovered in the 1980s. See [8] for information in this direction.

10 Further Reading

During the course of this article, I have mentioned some of my favorite articles and books about number theory, especially those that touch on Fermat's Last Theorem. Here are a few more references that I have not yet had occasion to cite. First, a summary of "elementary" approaches to Fermat's Last Theorem is provided by P. Ribenboim in his book [20]. Secondly, an interesting discussion of elliptic curves and modular forms is contained in A. van der Poorten's book [16]. Next, the recent "diary" by C. J. Mozzochi [15] contains photos of the mathematicians who participated in the proof of Fermat's Last Theorem, along with detailed descriptions of lectures and other events that are associated strongly with the proof. Finally, several accounts of the details of the proof of Fermat's Last Theorem have been written for professional mathematicians [7], [18], [6]. What is missing from the literature, at least so far, is an extended account of the proof that is accessible to a scientifically literate lay reader and does justice to the mathematics behind the proof.

Acknowledgments

The author thanks the University of Singapore and the Isaac Newton Institute for Mathematical Sciences at Cambridge for organizing this conference and for their kind hospitality in Singapore. He was partially supported by the US National Science Foundation during the preparation of this article.

References

1. M. Aigner and G. M. Ziegler, *Proofs from the Book.* New York-Berlin-Heidelberg: Springer-Verlag, 1998.
2. Z. I. Borevich and I. R. Shafarevich, *Number Theory.* New York: Academic Press, 1966.
3. C. Breuil, B. Conrad, F. Diamond, and R. Taylor, *On the modularity of elliptic curves over* **Q**. To appear.
4. J. Buhler, R Crandall, R. Ernvall, and T. Metsänkylä, *Irregular primes and cyclotomic invariants to four million*, Math. Comp. **61** (1993), 151–153.
5. B. Conrad, F. Diamond, and R. Taylor, *Modularity of certain potentially Barsotti-Tate Galois representations*, J. Amer. Math. Soc. **12** (1999), 521–567.
6. G. Cornell, J. H. Silverman and G. Stevens, eds. *Modular forms and Fermat's last theorem*, Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995. Berlin-Heidelberg-New York: Springer-Verlag, 1997.
7. H. Darmon, F. Diamond and R. L. Taylor, *Fermat's last theorem.* In "Elliptic curves, modular forms & Fermat's last theorem," Proceedings of the Conference on Elliptic Curves and Modular Forms held at the Chinese University of Hong Kong, Hong Kong, December 18–21, 1993, J. Coates and S. T. Yau, eds., second edition. Cambridge, MA: International Press, 1997.
8. H. Darmon and A. Granville, *On the equations* $z^m = F(x, y)$ *and* $Ax^p + By^q = Cz^r$, Bull. London Math. Soc. **27** (1995), 513–543.
9. H. Darmon and L. Merel, *Winding quotients and some variants of Fermat's last theorem*, J. Reine Angew. Math. **490** (1997), 81–100.
10. F. Diamond, *On deformation rings and Hecke rings*, Ann. of Math. (2) **144** (1996), 137–166.
11. H. M. Edwards, *Fermat's Last Theorem: a genetic introduction to modern number theory.* Graduate Texts in Mathematics, volume 50. New York-Berlin-Heidelberg: Springer-Verlag, 1977.

12. N. Elkies, *On $A^4 + B^4 + C^4 = D^4$*, Math. Comp. **51** (1988), 825–835.
13. Y. Hellegouarch, *Invitation aux mathématiques de Fermat-Wiles.* Paris: Masson, 1997.
14. K. F. Ireland and M. I. Rosen, *A Classical Introduction to Modern Number Theory*, section edition. Graduate Texts in Mathematics, volume 84. New York-Berlin-Heidelberg: Springer-Verlag, 1990.
15. C. J. Mozzochi, *The Fermat Diary.* Providence: American Math. Soc., 2000.
16. A. van der Poorten, *Notes on Fermat's Last Theorem.* New York: John Wiley & Sons., 1996.
17. J. Lynch and S. Singh, *The Proof.* A television documentary written and produced by John Lynch and directed by Simon Singh. See `http://www.pbs.org/wgbh/nova/proof/` for more information, and for a transcript.
18. V. K. Murty, ed., *Seminar on Fermat's Last Theorem*, Papers from the seminar held at the Fields Institute for Research in Mathematical Sciences, Toronto, Ontario, 1993–1994. Providence, American Math. Soc., 1995.
19. P. Ribenboim, *13 Lectures on Fermat's Last Theorem.* New York-Berlin-Heidelberg: Springer-Verlag, 1979.
20. P. Ribenboim, *Fermat's Last Theorem for Amateurs.* New York-Berlin-Heidelberg: Springer-Verlag, 1999.
21. K. A. Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ *arising from modular forms*, Invent. Math. **100** (1990), 431–476.
22. K. A. Ribet, *From the Taniyama-Shimura conjecture to Fermat's last theorem*, Ann. Fac. Sci. Toulouse Math. (5) **11** (1990), 116–139.
23. K. A. Ribet, *Galois representations and modular forms*, Bull. Amer. Math. Soc. (N.S.) **32** (1995), 375–402.
24. K. A. Ribet *On the equation $a^p + 2^\alpha b^p + c^p = 0$*, Acta Arithmetica **79** (1997) 7–16.
25. K. A. Ribet and B. Hayes, *Fermat's Last Theorem and modern arithmetic*, American Scientist (March–April, 1994), 144–156.
26. S. Singh, *Fermat's Enigma: The epic quest to solve the world's greatest mathematical problem, with a foreword by John Lynch.* New York: Walker and Co., 1997.
27. S. Singh and K. A. Ribet, *Fermat's Last Stand*, Scientific American **227** (1997), 68–73.
28. H. Stark, *On the "gap" in a theorem of Heegner,* J. Number Theory **1** (1969), 16–27.
29. R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke alge-*

bras, Annals of Math. **141** (1995), 553–572.

30. A. Weil, *Number Theory: an approach through history, From Hammurapi to Legendre.* Boston, Mass: Birkhäuser Boston 1984.
31. H. Weyl, *Algebraic Theory of Numbers.* Annals of Mathematics Studies, volume 1. Princeton: Princeton University Press 1940.
32. A. Wiles, *Modular elliptic curves and Fermat's Last Theorem,* Annals of Math. **141** (1995), 443–551.