Computational Number Theory

C. Pomerance

1 Introduction

Historically, computation has been a driving force in the development of mathematics. To help measure the sizes of their fields, the Egyptians invented geometry. To help predict the positions of the planets, the Greeks invented trigonometry. Algebra was invented to deal with equations that arose when mathematics was used to model the world. The list goes on, and it is not just historical. If anything, computation is more important than ever. Much of modern technology rests on algorithms that compute quickly: examples range from the WAVELETS that allow CAT scans, to the numerical extrapolation of extremely complex systems in order to predict weather and global warming, and to the combinatorial algorithms that lie behind Internet search engines (see Section **??** of THE MATHEMATICS OF ALGORITHM DESIGN).

In pure mathematics we also compute, and many of our great theorems and conjectures are, at root, motivated by computational experience. It is said that GAUSS, who was an excellent computationalist, needed only to work out a concrete example or two to discover, and then prove, the underlying theorem. While some branches of pure mathematics have perhaps lost contact with their computational origins, the advent of cheap computational power and convenient mathematical software has helped to reverse this trend.

One mathematical area where the new emphasis on computation can be clearly felt is number theory, and that is the main topic of this article. A prescient call-to-arms was issued by Gauss as long ago as 1801:

> The problem of distinguishing prime numbers from composite numbers, and of resolving the latter into their prime factors, is known to be one of the most important and useful in arithmetic. It has engaged the industry and wisdom of ancient and modern geometers to such an extent that it would be superfluous to discuss the problem at length. Nevertheless we must confess that all methods that have been proposed thus far are either restricted to very special cases or are so laborious and difficult that even for numbers that do not exceed the limits of tables constructed by estimable men, they try the patience of even the practiced calculator. And these methods do not apply at all to larger numbers... Further, the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated.

Factorization into primes is a very basic issue in number theory, but essentially all branches of number theory have a computational component. And in some areas there is such a robust computational literature that we discuss the algorithms involved as mathematically interesting objects in their own right. In this article we will briefly present a few examples of the computational spirit: in analytic number theory (the distribution of primes and the Riemann hypothesis); in Diophantine equations (Fermat's last theorem and the *abc* conjecture); and in elementary number theory (primality and factorization). A secondary theme that we shall explore is the strong and constructive interplay between computation, heuristic reasoning, and conjecture.

2 Distinguishing Prime Numbers from Composite Numbers

The problem is simple to state. Given an integer $n > 1$, decide if n is prime or composite. And we all know an algorithm. Divide n by each positive integer in turn. Either we find a proper factor, in which case we know that n is composite, or we do not, in which case we know that n is prime. For example, take $n = 269$. It is odd, so it has no even divisors. It is not a multiple of 3, so it has no divisor which is a multiple of 3. Continuing, we rule out 5, 7, 11, and 13. The next possibility, 17, has a square that is greater than 269, which means that if 269 were a multiple of 17, then it would also have to be a multiple of some number less than 17. Since we have ruled that out, we can stop our trial division at 13 and conclude that 269 is prime. (If we were actually carrying out the algorithm, we might try dividing 269 by 17, in which case we would discover that $269 = 15 \times 17 + 14$. At that point we would notice that the quotient, 15, is less than 17, which is what would tell us that 17^2 was greater than 269. Then we could stop.) In general, since a composite number n

has a proper factor d with $d \leqslant \sqrt{n}$, one can give up on the trial dividing once one passes $\sqrt{n}$, at which point we know that n is prime.

This straightforward method is excellent for mental computation with small numbers, and for machine computation for somewhat larger numbers. But it scales poorly, in that if you double the number of digits of n, then the time for the worst case is squared; it is therefore an "exponential time" algorithm. One might tolerate such an algorithm for 20-digit inputs, but think how long it would take to establish the primality of a 40-digit number! And you can forget about numbers with hundreds or thousands of digits. The issue of how the running time of an algorithm scales when one goes to larger inputs is absolutely paramount in measuring one algorithm against another. In contrast to the exponential time it takes to use trial division to recognize primes, consider the problem of multiplying two numbers. The school method of multiplication is to take each digit of one number in turn and multiply it by the other number, forming a parallelogram array. One then performs an addition to obtain the answer. If you now double the number of digits in each number, then the parallelogram becomes twice as large in each dimension, so the running time grows by a factor of about 4. Multiplication of two numbers is an example of a "polynomial time" algorithm; its runtime scales by a constant factor when the input length is doubled.

One might then rephrase Gauss's call to arms as follows. Is there a polynomial time algorithm that distinguishes prime numbers from composite numbers? Is there a polynomial time algorithm that can produce a nontrivial factor of a composite number? It might not be apparent at this point that these are two different questions, since trial division does both. We will see, though, that it is convenient to separate them, as did Gauss.

Let us focus on recognizing primes. What we would like is a simply computed criterion that primes satisfy and composites do not, or vice versa. An old theorem of Wilson might just fit the bill. Note that $6! = 720$, which is just one less than a multiple of seven. Wilson's theorem asserts that if n is prime, then $(n-1)! \equiv -1 \pmod{n}$. (The meaning of this and similar statements is explained in MODULAR ARITHMETIC.) This cannot hold when n is composite, for if p is a prime factor of n and is smaller than n, then it is a factor of $(n-1)!$, so it cannot possibly be a factor of $(n-1)!+1$. Thus, we have an ironclad criterion for primality. However, the Wilson criterion does not meet the standard of being simply computed, since we know no especially rapid way of computing factorials modulo another number. For example, Wilson predicts that $268! \equiv -1 \pmod{269}$, as we have already seen that 269 is prime. But if we did not know this already, how in the world could we quickly find the remainder when 268! is divided by 269? We can work out the product 268! one factor at a time, but this would take many more steps than trying divisors up to 17. It is hard to prove that something *cannot* be done, and in fact there is no theorem that says we cannot compute $a! \bmod b$ in polynomial time. We do know some ways of speeding up the computation over the totally naive method, but all methods known so far take exponential time. So, Wilson's theorem initially seems promising, but in fact it is no help at all unless we can find a fast way to compute $a! \bmod b$.

How about FERMAT'S LITTLE THEOREM? Note that $2^7 = 128$, which is 2 more than a multiple of 7. Or take $3^5 = 243$, which is 3 mod 5. Fermat's little theorem tells us that if n is prime and a is any integer, then $a^n \equiv a \pmod{n}$. If computing a large factorial modulo n is hard, perhaps it is also hard to compute a large power modulo n.

It cannot hurt to try it out for some moderate example to see if any ideas pop up. Take $a = 2$ and $n = 91$, so that we are trying to compute $2^{91} \bmod 91$. A powerful idea in mathematics is that of reduction. Can we reduce this computational problem to a smaller one? Notice that if we had already computed $2^{45} \bmod 91$, obtaining a remainder r_1, say, then $2^{91} \equiv 2r_1^2 \pmod{91}$. That is, it is just a short additional calculation to get to our goal, yet the power 45 is only half as big. How to continue is clear: we further reduce to the exponent 22, which is less than half of 45. If $2^{22} \bmod 91 = r_2$, then $2^{45} \equiv 2r_2^2 \pmod{91}$. And of course 2^{22} is the square of 2^{11}, and so on. It is not so hard to "automate" this procedure: the exponent sequence

$$1,\ 2,\ 5,\ 11,\ 22,\ 45,\ 91$$

can be read directly from the binary (base 2) representation of 91 as 1011011, since the above sequence in binary is

$$1,\ 10,\ 101,\ 1011,\ 10110,\ 101101,\ 1011011.$$

These are the initial strings from the left of 1011011. And it is plain that the transition from one term to the next is either the double or the double plus 1.

This procedure scales nicely. When the number of digits of n is doubled, so is the sequence of exponents, and the time it takes to get from one exponent to the next, being a modular multiplication, is multiplied by 4. (As with naive multiplication, naive divide-with-remainder also takes four times as long when the size of the problem is doubled.) Thus, the overall time is multiplied by 8, yielding a polynomial time method. We call this the "powermod" algorithm.

So, let us try to illustrate Fermat's little theorem, taking $a = 2$ and $n = 91$. Our sequence of powers is

$$2^1 \equiv 2, \quad 2^2 \equiv 4, \quad 2^5 \equiv 32, \quad 2^{11} \equiv 46,$$
$$2^{22} \equiv 23, \quad 2^{45} \equiv 57, \quad 2^{91} \equiv 37,$$

where each congruence is modulo 91, and each term in the sequence is found by squaring the prior one mod 91 or squaring and multiplying by 2 mod 91.

Wait a second: does Fermat's little theorem not say that we are supposed to get 2 for the final residue? Well, yes, but this is guaranteed only if n is prime. And as you have probably already noticed, 91 is composite. In fact, the computation proves this.

Quite remarkably, here is an example of a computation that proves that n is composite, yet it does not reveal any nontrivial factorization!

You are invited to try out the powermod algorithm as above, but to change the base of the power from 2 to 3. The answer you should come to is that $3^{91} \equiv 3 \pmod{91}$: that is, the congruence for Fermat's little theorem holds. Since you already know that 91 is composite, I am sure you would not jump to the false conclusion that it is prime! So, as it stands, Fermat's little theorem can be used to sometimes recognize composites, but it cannot be used to recognize primes.

There are two interesting further points to be made regarding Fermat's little theorem. First, on the negative side, there are some composites, such as $n = 561$, where the Fermat congruence holds for *every* integer a. These numbers n are called *Carmichael numbers*, and unfortunately (from the point of view of testing primality) there are infinitely many of them, a result due to Alford, Granville, and me. But, on the positive side, if one were to choose randomly among all pairs a, n for which $a^n \equiv a \pmod{n}$, with $a < n$ and n bounded by a large number x, almost certainly (as x grows) you would choose a pair with n prime, a result of Erdős and myself.

It is possible to combine Fermat's little theorem with another elementary property of (odd) prime numbers. If n is an odd prime, there are exactly two solutions to the congruence $x^2 \equiv 1 \pmod{n}$, namely ± 1. Actually, some composites have this property as well, but composites divisible by two different odd primes do not.

Now let us suppose that n is an odd number and that we wish to determine whether it is prime. Suppose that we pick some number a with $1 \leqslant a \leqslant n-1$ and discover that $a^{n-1} \equiv 1 \pmod{n}$. If we set $x = a^{(n-1)/2}$, then $x^2 = a^{n-1} \equiv 1 \pmod{n}$; so, by the simple property of primes just mentioned, if n is prime, then x must be ± 1. Therefore, if we calculate $a^{(n-1)/2}$ and discover that it is not congruent to $\pm 1 \pmod{n}$, then n must be composite.

Let us try this idea with $a = 2$, $n = 561$. We know already that $2^{560} \equiv 1 \pmod{561}$, so what is 2^{280} mod 561? This too turns out to be 1, so we have not shown that 561 is composite. However, we can go further, since now we know that 2^{140} is also a square root of 1 and computing this we find that $2^{140} \equiv 67 \pmod{561}$. So now we have found a square root of 1 that is not ± 1, which proves that 561 is composite. (Of course, for this particular number, it is obviously divisible by 3, so there was not really any mystery about whether it was prime or composite. But the method can be used in much less obvious cases.) In practice, there is no need to backtrack from a higher exponent to a smaller one. Indeed, in order to calculate $2^{560} \pmod{561}$ by the efficient method outlined earlier, one calculates the numbers 2^{140} and 2^{280} along the way, so that this generalization of the earlier test is both quicker and stronger.

Here is the general principle that we have illustrated. Suppose that n is an odd prime and let a be an integer not divisible by n. Write $n - 1 = 2^s t$, where t is odd. Then

$$\text{either } a^t \equiv 1 \pmod{n} \quad \text{or} \quad a^{2^i t} \equiv -1 \pmod{n}$$

for some $i = 0, 1, \dots, s-1$. Call this the *strong Fermat congruence*. The wonderful thing here is that there is no analogue of a Carmichael number—as proved independently by Monier and Rabin. They showed that if n is an odd composite, then the strong Fermat congruence fails for at least three-quarters of the choices for a with $1 \leqslant a \leqslant n-1$.

If you want only to be able to distinguish between primes and composites in practice, and you do not insist on proof, then you have read enough. Namely, given a large odd number n, choose 20 values of a at random from $[1, n-1]$, and begin trying to verify the strong Fermat congruence with these bases a. If it should ever fail, you may stop: the number n must be composite. And if the strong Fermat congruence holds, we might surmise that n is actually prime. Indeed, if n were composite, the Monier–Rabin theorem says that the chance that the strong Fermat congruence would hold for 20 random bases is at most 4^{-20}, which is less than one chance in a trillion. Thus we have a remarkable *probabilistic* test for primality. If it tells us that n is composite, then we know for sure that n is composite; if it tells us that n is prime, then the chances that n is not prime are so small as to be more or less negligible.

If three-quarters of the numbers a in $[1, n-1]$ provide the key to an easily checkable proof that the odd composite number n is indeed composite, surely it should not be so hard to find just one! How about checking small numbers a, in order, until one is found? Excellent, but when do we stop? Let us think about this for a moment. We have given up the power of randomness and are forcing ourselves to choose sequentially among small numbers for the trial bases a. Can we argue heuristically that they continue to behave as if they were random choices? Well, there *are* some connections among them. For example, if taking $a = 2$ does not result in a proof that n is composite, then neither will taking any power of 2. It is theoretically possible for 2 and 3 not to give proofs that n is composite but for 6 to work just fine, but this turns out not to be very common. So let us amend the heuristic and assume that we have independence for *prime* values of a. Up to $\log n \log\log n$ there are about $\log n$ primes (via the PRIME NUMBER THEOREM discussed later in this article); so, heuristically, the probability that n is composite, but that none of these primes help us to prove it, is about $4^{-\log n} < n^{-4/3}$. Since the infinite sum $\sum n^{-4/3}$ converges, perhaps a stopping point of $\log n \log\log n$ is sufficient, at least for large n.

Miller was able to prove the slightly weaker result that a stopping point of $c(\log n)^2$ is adequate, but his proof assumes a generalization of the RIEMANN HYPOTHESIS. (We discuss the Riemann hypothesis below; the generalization that Miller assumes is beyond the scope of this article.) In further work, Bach was able to show that we may take $c = 2$ in this last result. Summarizing, if this generalized Riemann hypothesis holds, and if the strong Fermat congruence holds for every positive integer $a \leqslant 2(\log n)^2$, then n is prime. So, provided that a famous unproved hypothesis in another field of mathematics is correct, one can decide in polynomial time, via a deterministic algorithm, whether n is prime or composite. (It has been tempting to *use* this conditional test, for if it should ever lie to you and tell you that a particular composite number is prime, then this failure—if you were able to detect it—would be a disproof of one of the most famous conjectures in mathematics. Perhaps this is not too disastrous a failure!)

After Miller's test in the 1970s, the question continually challenging us was whether it is possible to test for primality in polynomial time without assuming unproved hypotheses. Recently, Agrawal et al. (2004) answered this question with a resounding yes. Their idea begins with a combination of the binomial theorem and Fermat's little theorem. Given an integer a, consider the polynomial $(x+a)^n$ and expand it in the usual way through the binomial theorem. Each intermediate term between the leading x^n and the trailing a^n has the coefficient $n!/(j!(n-j)!)$ for some j between 1 and $n-1$. If n is prime, then this coefficient, which is an integer, is divisible by n because n appears as a factor in the numerator that is not cancelled by any factors in the denominator. That is, the coefficient is $0 \pmod{n}$. For example, $(x+1)^7$ is equal to

$$x^7 + 7x^6 + 21x^5 + 35x^4 + 35x^3 + 21x^2 + 7x + 1,$$

and we see each internal coefficient is a multiple of 7. Thus, we have $(x+1)^7 \equiv x^7 + 1 \pmod{7}$. (Two polynomials are congruent mod n if corresponding coefficients are congruent mod n.) In general, if n is prime and a is any integer, then via this binomial-theorem idea and Fermat's little theorem we have

$$(x+a)^n \equiv x^n + a^n \equiv x^n + a \pmod{n}.$$

It is an easy exercise to show that this congruence in the simple case $a = 1$ is actually equivalent to primality. But as with the Wilson criterion we know no way of quickly verifying that all these coefficients are indeed divisible by n.

However, one can do more with polynomials than raise them to powers. We can also divide one polynomial by another to find a quotient and a remainder, just as we do with integers. It makes sense, for example, to say that $g(x) \equiv h(x) \pmod{f(x)}$, meaning that $g(x)$ and $h(x)$ leave the same remainder when divided by $f(x)$. We will write $g(x) \equiv h(x)$ $(\text{mod } n, f(x))$ if the remainders upon division by $f(x)$ are congruent mod n. As with the powermod algorithm for integer congruences, we can quickly compute $g(x)^n \pmod{n, f(x)}$, provided the degree of $f(x)$ is not too big. This is exactly what Agrawal et al. propose. They have an auxiliary polynomial $f(x)$ of not-too-high degree such that, if

$$(x+a)^n \equiv x^n + a \ (\operatorname{mod} n, f(x))$$

for each $a = 1, 2, \dots, B$, for a not-too-high bound B, then n must be in a set that contains the primes and certain composites that are easily recognized as composites. (Not all composites are hard to recognize as such, e.g., any number with a small prime factor is easy to recognize.) These ideas put together form the primality test of Agrawal et al. To give the argument in full detail one has to specify the auxiliary polynomial $f(x)$ that is used and what the bound B is, and one has to prove rigorously that it is exactly the primes which pass the test.

Agrawal et al. (2004) show that the auxiliary polynomial $f(x)$ can be taken to be the beautifully simple $x^r - 1$, with an elementary upper bound for r of about $(\log n)^5$. Doing this leads to a time bound of about $(\log n)^{10.5}$ for the algorithm. Using a numerically ineffective tool, they bring the time bound down to $(\log n)^{7.5}$. Recently, Lenstra and Pomerance (forthcoming) presented a not-so-simple but numerically effective method of bringing the exponent on $\log n$ down to 6. We did this by expanding the set of polynomials used beyond those of the form $x^r - 1$: in particular we used polynomials that are related to Gauss's famous algorithm for construction of certain regular n-gons with straightedge and compass. It was indeed satisfying to us to bring in a famous tool of Gauss to say something about his problem of distinguishing prime numbers from composite numbers.

Are the new polynomial-time primality tests good in practice? So far, the answer is no, the competition is just too tough. For example, using the arithmetic of elliptic curves we can come up with bona fide proofs of primality for huge numbers. This algorithm is conjectured to run in polynomial time but we have not even proved that it always terminates. If, at the end of the day, or in this case the end of the run, we have a legitimate proof, then perhaps we can tolerate the situation of not being sure that it would work out when we started! The method, pioneered by Atkin and Morain, has recently proved the primality of a number that has over 20 000 decimal digits, and is not of some special form such as $2^n - 1$ that makes testing for primality easier. The record for the new breed of polynomial-time tests is a measly 300 digits.

For numbers of certain special forms there are much faster primality tests. Mersenne primes comprise the most famous of these forms; these are primes that are 1 less than a power of 2. It is suspected that there are infinitely many examples, but we seem to be very far from a proof of this. Just 43 Mersenne primes are known, the record example being $2^{30\,402\,457}-1$, a prime with more than 9.15 million decimal digits.

For much more on primality testing, and for references to various other sources, see Crandall and Pomerance (2005).

3 Factoring Composite Numbers

Compared with what we know about testing primality, our ability to factor large numbers is still in the dark ages. In fact this imbalance between the two problems forms the bulwark for the security of electronic commerce on the Internet. (See public-key cryptography for an account of why.) This is a very important application of mathematics, but also an odd one, and not something to brag about, since it depends on the inability of mathematicians to efficiently solve a basic problem!

Nevertheless, we do have our tricks. Part of the landscape is Euclid's algorithm for computing the greatest common divisor (GCD) of two numbers. One might naively think that, to find the GCD of two positive integers m and n, one should find all of their divisors and pick the largest one common to the two. But Euclid's algorithm is much more efficient: the number of arithmetic steps is bounded by the logarithm of the smaller number, so not only does it run in polynomial time, it is in fact quite speedy. (See The Euclidean algorithm and continued fractions for more details.)

So, if we can build up a special number m that may be likely to have a nontrivial factor in common with n, we can use Euclid's algorithm to discover this factor. For example, Pollard and Strassen (independently) used this idea, together with fast subroutines for multiplication and polynomial evaluation, to enhance the trial division method discussed in the last section. Somewhat miraculously, one can take the integers up to $n^{1/2}$, break them into $n^{1/4}$ subintervals of length $n^{1/4}$, and for each subinterval calculate the GCD of n with the product of all the integers in the subinterval, spending only about $n^{1/4}$ elementary steps in total. If n is composite, then at least one GCD will be larger than 1, and then a search over the first such subinterval will locate a nontrivial factor of n. To date, this algorithm is the fastest rigorous and deterministic method of factoring that we know.

Most practical factoring algorithms are based on unproved but reasonable-seeming hypotheses about the natural numbers. Although we may not know how to prove rigorously that these methods will always produce a factorization, or do so quickly, in practice they do. This situation resembles the experimental sciences, where hypotheses are tested against experiments. Our experience with certain factoring algorithms is now so overwhelming that a scientist might claim that a physical law is involved. As mathematicians, we still search for proof, but fortunately the numbers we factor do not feel the need to wait for us.

I often mention a contest problem from my high school years: factor 8051. The trick is to notice that $8051 = 90^2 - 7^2 = (90-7)(90+7)$, from which the factorization $83 \cdot 97$ can be read off. In fact every odd composite can be factored as the difference of two squares, an idea that goes back to Fermat. Indeed, if n has the nontrivial factorization ab, then let $u = \frac{1}{2}(a+b)$ and $v = \frac{1}{2}(a-b)$, so that $n = u^2 - v^2$, and $a = u+v$, $b = u - v$. This method works very well if n has a divisor very close to $n^{1/2}$, as $n = 8051$ does, but in the worst case, the Fermat method is slower than trial division.

My quadratic sieve method (which follows work of Kraitchik, Brillhart–Morrison, and Schroeppel) tries to efficiently extend Fermat's idea to all odd composites. For example, take $n = 1649$. We start just above $n^{1/2}$ with $j = 41$, and consider the numbers $j^2 - 1649$. As j runs, we will eventually hit a value where $j^2 - 1649$ is a square, and so be able to use Fermat's method. Let's try it:

$$\begin{aligned} 41^2 - 1649 &= 32, \\ 42^2 - 1649 &= 115, \\ 43^2 - 1649 &= 200, \\ &\vdots \end{aligned}$$

Well, no squares yet, which is not surprising, since the Fermat method is often very poor. But wait, do the first and third lines not multiply together to give a square? Yes they do, $32 \cdot 200 = 80^2$. So, multiplying the first and third lines, and treating them as congruences mod 1649, we have

$$(41 \cdot 43)^2 \equiv 80^2 \pmod{1649}.$$

That is, we have a pair u, v with $u^2 \equiv v^2 \pmod{1649}$. This is not quite the same as having $u^2 - v^2 = 1649$, but we do have 1649 a divisor of $u^2 - v^2 = (u-v)(u+v)$. Now maybe 1649 divides one of these factors, but if it does not, then it is split between them, and so a computation of the GCD of $u - v$ (or $u + v$) with 1649 will reveal a proper factor. Now $v = 80$ and $u = 41 \cdot 43 \equiv 114 \pmod{1649}$, and so we see instantly that $u \not\equiv \pm v \pmod{1649}$, so we are in business. The GCD of $114 - 80 = 34$ with 1649 is 17. Dividing, we see that $1649 = 17 \cdot 97$, and we are done.

Can we generalize this? In trying to factor $n = 1649$ we considered consecutive values of the quadratic polynomial $f(j) = j^2 - n$ for j starting just above $\sqrt{n}$, and viewed these as congruences $j^2 \equiv f(j) \pmod{n}$. Then we found a set $\mathcal{M}$ of numbers j with $\prod_{j\in\mathcal{M}} f(j)$ equal to a square, say v^2. We then let $u = \prod_{j\in\mathcal{M}} j$, so that $u^2 \equiv v^2 \pmod{n}$. Since $u \not\equiv \pm v \pmod{n}$, we could split n via the GCD of $u - v$ and n.

There is another lesson that we can learn from our small example with $n = 1649$. We used 32 and 200 to form our square, but we ignored 115. If we had thought about it, we might have noticed from the start that 32 and 200 were more likely to be useful than 115. The reason is that that 32 and 200 are *smooth* numbers (meaning that they have only small prime factors), while 115 is not smooth, having the relatively large prime factor 23. Say you have $k+1$ positive integers that involve in their prime factorizations only the first k primes. It is an easy theorem that some nonempty subset of these numbers has product a square. The proof has us associate with each of these numbers, which can be written in the form $p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}$, an *exponent vector*

$(a_1, a_2, \ldots, a_k)$. Since squares are detected by all even exponents, we really only care whether the exponents a_i are odd or even. Thus, we think of these vectors as having coordinates 0 and 1, and when we add them (which corresponds to multiplying the underlying numbers), we do so mod 2. Since we have $k+1$ vectors, each with only k coordinates, an easy matrix calculation leads quickly to a nonempty subset that adds up to the 0-vector. The product of the corresponding integers is then a square.

In our toy example with $n = 1649$, the first and third numbers, which are $32 = 2^5 3^0 5^0$ and $200 = 2^3 3^0 5^2$, have exponent vectors $(5, 0, 0)$ and $(3, 0, 2)$, which reduce to $(1, 0, 0)$ and $(1, 0, 0)$, so we see that the sum of them is $(0, 0, 0)$, which indicates that we have a square. We were lucky that we could make do with just two vectors, instead of the four that the above argument shows would be sufficient.

In general with the quadratic sieve, one finds smooth numbers in the sequence $j^2 - n$, forms the exponent vectors mod 2, and then uses a matrix to find a nonempty subset which adds up to the 0-vector, which then corresponds to a set $\mathcal{M}$ where $\prod_{j \in \mathcal{M}} f(j)$ is a square.

In addition, the "sieve" in the quadratic sieve comes in with the search for smooth values of $f(j) = j^2 - n$. These numbers are the consecutive values of a (quadratic) polynomial, so those divisible by a given prime can be found in regular places in the sequence. For example, in our illustration, $j^2 - 1649$ is divisible by 5 precisely when $j \equiv 2$ or 3 (mod 5). A sieve, very much like the sieve of Eratosthenes, can then be used to efficiently find the special numbers j where $j^2 - n$ is smooth. A key issue though is how smooth a value $f(j)$ has to be for us to decide to accept it. If we choose a smaller bound for the primes involved, we do not have to find all that many of them to use the matrix method. But such very smooth values might be very rare. If we use a larger bound for the primes involved, then smooth values of $f(j)$ may be more common, but we will need many of them. Somewhere between smaller and larger is just right! In order to make the choice, it would help to know how frequently values of an irreducible quadratic polynomial are smooth. Unfortunately, we do not have a theorem that tells us, but we can still make a good choice by assuming that this frequency is about that for a random number of the same size, an assumption that is probably correct even if it is hard to prove.

Finally, note that if the final GCD yields only a trivial factor with n, one can continue just a bit longer and find more linear dependencies, each with a fresh chance at splitting n.

These thoughts lead us to a time bound of about

$$\exp(\sqrt{\log n \, \log\log n}\,)$$

for the quadratic sieve to factor n. Instead of being exponential in the number of digits of n, as with trial division, this is exponential in about the square root of the number of digits of n. This is certainly a huge improvement, but it is still a far cry from polynomial time.

Lenstra and I actually have a rigorous random factoring method with the same time complexity as that above for the quadratic sieve. (It is random in the sense that a coin is flipped at various junctures, and decisions on what to do next depend on the outcomes of these flips. Through this process, we expect to get a bona fide factorization within the advertised time bound.) However, the method is not so computer practical, and if you had to choose in practice between the two, then you should go with the nonrigorous quadratic sieve. A triumph for the quadratic sieve was the 1994 factorization of the 129-digit RSA cryptographic challenge first published in Martin Gardner's column in *Scientific American* in 1977.

The *number field sieve*, which is another sieve-based factoring algorithm, was discovered in the late 1980s by Pollard for integers close to powers, and later developed by Buhler, Lenstra and me for general integers. The method is similar in spirit to the quadratic sieve, but assembles its squares from the product of certain sets of algebraic integers. The number field sieve has a conjectured time complexity of the type

$$\exp(c(\log n)^{1/3}(\log\log n)^{2/3}),$$

for a value of c slightly below 2. For composite numbers beyond 100 digits or so that have no small prime factor, it is the method of choice, with the current record being 200 decimal digits.

The sieve-based factorization methods share the property that if you use them, then all composite numbers of about the same size are equally hard to factor. For instance, factoring n will be about as difficult if n is a product of five primes each roughly near the fifth root of n as it will be if n is a product of two primes roughly near the square root of n. This is quite unlike trial division, which is happiest when there is a small prime factor. We will now describe a famous factorization method due to Lenstra that detects small prime factors before large ones, and beyond baby cases is much superior to trial dividing. This is his *elliptic curve method*.

Just as the quadratic sieve searches for a number m with a nontrivial GCD with n, so does the elliptic curve method. But where the quadratic sieve (and the number field sieve) painstakingly build up to a successful m from many small successes, the elliptic curve method hopes to hit upon m with essentially one lucky choice.

Choosing random numbers m and testing their GCD with n can also have instant success, but you can well imagine that if n has no small prime factors, then the expected time for success would be enormous. Instead, the elliptic curve method involves considerably more cleverness.

Consider first the "$p-1$ method" of Pollard. Suppose you have a number n you wish to factor and a certain large number k. Unbeknownst to you, n has a prime factor p with $p-1$ a divisor of k, and another prime factor q with $q-1$ not a divisor of k. You can use this imbalance to split n. First of all, by Fermat's little theorem there are many numbers u with $u^k \equiv 1 \pmod{p}$ and $u^k \not\equiv 1 \pmod{q}$. Say you have one of these, and let m be $u^k - 1$ reduced mod n. Then the GCD of m and n is a nontrivial factor of n; it is divisible by p but not by q. Pollard suggests taking k as the least common multiple of the integers to some moderate bound so that it has many divisors and perhaps a decent chance that it is divisible by $p-1$. The best case of Pollard's method is when n has a prime factor p with $p-1$ smooth (has all small prime factors—see the quadratic sieve discussion above). But if n has no prime factors p with $p-1$ smooth, Pollard's method fares poorly.

What is going on here is that corresponding to the prime p we have the multiplicative group of the $p-1$ nonzero residues mod p. Furthermore, when doing arithmetic mod n with numbers relatively prime to n, we are, whether we realize it or not, doing arithmetic in this group. We are exploiting the fact that u^k is the group identity mod p, but not mod q.

Lenstra had the brilliant idea of using the Pollard method in the context of elliptic curve groups. There are many elliptic curve groups associated with the prime p, and therefore many chances to hit upon one where the number of elements is smooth. Of great importance here are theorems of Hasse and Deuring. An elliptic curve mod p (for $p > 3$) can be taken as the set of solutions to the congruence $y^2 \equiv x^3 + ax + b \pmod{p}$, for given integers a, b with the property that $x^3 + ax + b$ does not have repeated roots mod p. There is one additional "point at infinity" thrown in (see below). A fairly simple addition law (but not as simple as adding coordinatewise!) makes the elliptic curve into a group, with the point at infinity as the identity. Hasse, in a result later generalized by Weil with his famous proof of the "Riemann hypothesis for curves," showed us that the number of elements in the elliptic curve group always lies between $p + 1 - 2\sqrt{p}$ and $p + 1 + 2\sqrt{p}$. And Deuring proved that every number in this range is indeed associated with some elliptic curve mod p.

Say we randomly choose integers x_1, y_1, a, and then choose b so that y_1^2 is congruent to $x_1^3 + ax_1 + b \pmod{n}$. This gives us the curve with coefficients a, b and a point $P = (x_1, y_1)$ on the curve. One can then mimic the Pollard strategy, with a number k as before with many divisors, and with the point P playing the role of u. Let kP denote the k-fold sum of P added to itself using elliptic curve addition. If kP is the point at infinity on the curve considered mod p (which it will be if the number of points on the curve is a divisor of k), but not on the curve considered mod q, then this gives us a number m whose GCD with n is divisible by p and not by q. We will have factored n.

To see where m comes from it is convenient to consider the curve projectively: we take solutions (x, y, z) of the congruence $y^2 z \equiv x^3 + axz^2 + bz^3 \pmod{p}$. The triple (cx, cy, cz) when $c \neq 0$ is considered to be the same as (x, y, z). The mysterious point at infinity is now demystified; it is just $(0, 1, 0)$. And our point P is $(x_1, y_1, 1)$. (This is the mod p version of classical projective geometry which is described in Section **??** of Some Fundamental Mathematical Definitions.) Say we work mod n and compute the point $kP = (x_k, y_k, z_k)$. Then the candidate for the number m is just z_k. Indeed, if kP is the point at infinity mod p, then $z_k \equiv 0 \pmod{p}$, and if it is not the point at infinity mod q, then $z_k \not\equiv 0 \pmod{q}$.

When Pollard's $p-1$ method fails, our only recourse is to raise k or give up. With the elliptic curve method, if things do not work for our randomly chosen curve, we can pick another. Corresponding to the hidden prime p in n, we are actually picking new elliptic curve groups mod p, and so gaining a fresh chance for the number of elements in the group to be smooth. The elliptic curve method has been quite successful in factoring numbers which have a prime factor up to about 50 decimal digits, and even occasionally somewhat larger primes have been discovered.

We conjecture that the expected time for the elliptic curve method to find the least prime factor p of n is about

$$\exp(\sqrt{2 \log p \, \log\log p}\,)$$

arithmetic operations mod n. What is holding us back from proving this conjecture is not lack of knowledge about elliptic curves, but rather lack of knowledge of the distribution of smooth numbers.

For more on these and other factorization methods, the reader is referred to Crandall and Pomerance (2005).

4 The Riemann Hypothesis and the Distribution of the Primes

As a teenager looking at a modest table of primes, Gauss conjectured that their frequency decays logarithmically and that $\mathrm{li}(x) = \int_2^x \mathrm{d}t/\log t$ should be a good approximation for $\pi(x)$, the number of primes between 1 and x. Sixty years later, Riemann showed how Gauss's conjecture can be proved if one assumes that the Riemann zeta function $\zeta(s) = \sum_n n^{-s}$ has no zeros in the complex half-plane where the real part of s is greater than $\frac{1}{2}$. The series for $\zeta(s)$ converges only for $\mathrm{Re}\, s > 1$, but it may be analytically continued to $\mathrm{Re}\, s > 0$, with a simple pole at $s = 1$. (For a brief description of the process of analytic continuation, see Section **??** of Some fundamental definitions.) This continuation may be seen quite concretely via the identity $\zeta(s) = s/(s-1) - s\int_1^\infty \{x\}x^{-s-1}\,\mathrm{d}x$, with $\{x\}$ the fractional part of x (so that $\{x\} = x - [x]$): note that this integral converges quite nicely in the half-plane $\mathrm{Re}\, s > 0$. In fact, via Riemann's functional equation mentioned below, $\zeta(s)$ can be continued to a meromorphic function in the whole complex plane, with the single pole at $s = 1$.

The assertion that $\zeta(s) \neq 0$ for $\mathrm{Re}\, s > \frac{1}{2}$ is known as the Riemann hypothesis; arguably it is the most famous unsolved problem in mathematics. Though Hadamard and de la Vallee Poussin were able in 1896 to prove (independently) a weak form of Gauss's conjecture known as the prime number theorem, the apparent breathtaking strength of the approximation $\mathrm{li}(x)$ to $\pi(x)$ is uncanny. For example, take $x = 10^{22}$. We have

$$\pi(10^{22}) = 201\,467\,286\,689\,315\,906\,290$$

exactly, and, to the nearest integer, we have

$$\mathrm{li}(10^{22}) \approx 201\,467\,286\,691\,248\,261\,497.$$

As you can plainly see, Gauss's guess is right on the money!

The numerical computation of $\mathrm{li}(x)$ is simple via numerical methods for integration, and it is directly obtainable in various mathematics computing packages. However, the computation of $\pi(10^{22})$ (due to Gourdon) is far from trivial. It would be far too laborious to count these approximately 2×10^{20} primes one by one, so how are they counted? In fact, we have various combinatorial tricks to count without listing everything. For example, one does not need to count one by one to see that there are exactly $2[10^{22}/6] + 1$ integers in the interval from 1 to 10^{22} that are relatively prime to 6. Rather one thinks of these numbers grouped in blocks of six, with two in each block coprime to 6. (The "+1" comes from the partial block at the end.) Building on early ideas of Meissel and Lehmer, Lagarias, Miller, and Odlyzko presented an elegant combinatorial method for computing $\pi(x)$ that takes about $x^{2/3}$ elementary steps. The method was refined by Deléglise and Rivat, and then Gourdon found a way to distribute the computation to many computers.

From work of von Koch, and later Schoenfeld, we know that the Riemann hypothesis is *equivalent* to the assertion that

$$|\pi(x) - \mathrm{li}(x)| < \sqrt{x}\log x \qquad (1)$$

for all $x \geqslant 3$ (see Crandall and Pomerance 2005, Exercise 1.37). Thus, the mammoth calculation of $\pi(10^{22})$ might be viewed as computational evidence for the Riemann hypothesis—in fact, if the count had turned out to violate (1), we would have had a disproof.

It may not be obvious what (1) has to do with the location of the zeros of $\zeta(s)$. To understand the connection, let us first dismiss the so-called "trivial" zeros, which occur at each negative even integer. The nontrivial zeros ρ are known to be infinite in number, and, as mentioned above, are conjectured to satisfy $\operatorname{Re}\rho \leqslant \frac{1}{2}$. There are certain symmetries among these zeros: indeed, if ρ is a zero, then so are $\bar{\rho}$, $1-\rho$, and $1-\bar{\rho}$. Therefore, the Riemann hypothesis is the assertion that every nontrivial zero has real part equal to $\frac{1}{2}$. (The symmetry with ρ and $1-\rho$, which follows from Riemann's functional equation $\zeta(1-s) = 2(2\pi)^{-s}\cos(\frac{1}{2}\pi s)\Gamma(s)\zeta(s)$, perhaps provides some heuristic support for the Riemann hypothesis.)

The connection to prime numbers begins with the fundamental theorem of arithmetic, which yields the identity

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_{p \text{ prime}} \sum_{j=0}^{\infty} p^{-js} = \prod_{p \text{ prime}} (1-p^{-s})^{-1},$$

a product that converges when $\operatorname{Re} s > 1$. Thus, taking the logarithmic derivative (that is, taking the logarithm of both sides and then differentiating), we have

$$\frac{\zeta'(s)}{\zeta(s)} = -\sum_{p \text{ prime}} \frac{\log p}{p^s - 1} = -\sum_{p \text{ prime}} \sum_{j=1}^{\infty} \frac{\log p}{p^{js}}.$$

That is, if we define $\Lambda(n)$ to be $\log p$ if $n = p^j$ for a prime p and an integer $j \geqslant 1$, and $\Lambda(n) = 0$ if n is not of this form, then we have the identity

$$\sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = -\frac{\zeta'(s)}{\zeta(s)}.$$

Through various relatively routine calculations, one can then relate the function

$$\psi(x) = \sum_{n \leqslant x} \Lambda(n)$$

to the residues at the poles of ζ'/ζ, which correspond to the zeros (and single pole) of ζ. In fact, as Riemann showed, we have the following beautiful formula:

$$\psi(x) = x - \sum_{\rho} \frac{x^{\rho}}{\rho} - \log(2\pi) - \tfrac{1}{2}\log(1-x^{-2})$$

if x itself is not a prime or prime power, and where the sum over the nontrivial zeros ρ of ζ is to be understood in the symmetric sense where we sum over those ρ with $|\operatorname{Im}\rho| < T$ and let $T \to \infty$. Through elementary manipulations, an understanding of the function $\psi(x)$ readily gives an equivalent understanding of $\pi(x)$, and it should be clear now that $\psi(x)$ is intimately connected to the nontrivial zeros ρ of ζ.

The function $\psi(x)$ defined above has a simple interpretation. It is the logarithm of the least common multiple of the integers in the interval $[1, x]$. As with (1) we have an elementary translation of the Riemann hypothesis: it is equivalent to the assertion that

$$|\psi(x) - x| < \sqrt{x}\log^2 x$$

for all $x \geqslant 3$. This inequality involves only the elementary concepts of least common multiple, natural logarithm, absolute value, and square root, yet it is equivalent to the Riemann hypothesis.

A number of nontrivial zeros ρ of $\zeta(s)$ have actually been calculated and it has been verified that they lie on the line $\operatorname{Re} s = \frac{1}{2}$. One might wonder how someone can computationally verify that a complex number ρ has $\operatorname{Re}\rho = \frac{1}{2}$. For example, suppose that we are carrying calculations to (an unrealistically large) 10^{10} significant digits, and suppose we come across a zero with real part $\frac{1}{2} + 10^{-10^{100}}$. It would be far beyond the precision of the calculation to be able to distinguish this number from $\frac{1}{2}$ itself. Nevertheless, we do have a method for seeing if particular zeros ρ satisfy $\operatorname{Re}\rho = \frac{1}{2}$. There are two ideas involved, one of which comes from elementary calculus. If we have a continuous real-valued function $f(x)$ defined on the real numbers, we can sometimes use the intermediate value theorem to count zeros. For example, say $f(1) > 0$, $f(1.7) < 0$, $f(2.3) > 0$. Then we know for sure that f has at least one zero between 1 and 1.7, and at least one zero between 1.7 and 2.3. If we know for other reasons that f has exactly two zeros, then we have accounted for both of them. To locate zeros of the complex function $\zeta(s)$, a real-valued function $g(t)$ is constructed with the property that $\zeta(\frac{1}{2} + \mathrm{i}t) = 0$ if and only if $g(t) = 0$. By looking at sign changes for $g(t)$ for $0 < t < T$, we can get a *lower bound* for the number of zeros ρ of ζ with $\operatorname{Re}\rho = \frac{1}{2}$ and $0 < \operatorname{Im}\rho < T$. In addition, we can use the so-called *argument principle* from complex analysis to count the *exact number* of zeros with $0 < \operatorname{Im}\rho < T$. If we are lucky and this exact count is equal to our lower bound, then we have accounted for all of ζ's zeros here, showing that they all have real part $\frac{1}{2}$ (and, in addition, that they are all simple zeros). If the counts did not match, it would not be a disproof of the Riemann hypothesis, but certainly it would indicate a region where we should be checking the data more closely. So far, whenever we have tried this approach, the counts have matched, though sometimes we have been forced to evaluate $g(t)$ at very closely spaced points.

The first few nontrivial zeros were computed by Riemann himself. The famous cryptographer and early computer scientist Turing also computed some zeta zeros. The current record for this kind of calculation is held by Gourdon, who has shown that the first 10^{13} zeta zeros with positive imaginary part all have real part equal to $\frac{1}{2}$, as predicted by Riemann. Gourdon's method is a modification of that pioneered by Odlyzko and Schönhage (1988), who ushered in the modern age of zeta-zero calculations.

Explicit zeta-function calculations can lead to highly useful explicit prime number estimates. If p_n is the nth prime, then the prime number theorem implies that $p_n \sim n \log n$ as $n \to \infty$. Actually, there is a secondary term of order $n \log \log n$, and so for all sufficiently large n, we have $p_n > n \log n$. By using explicit zeta estimates, Rosser was able to put a numerical bound on the "sufficiently large" in this statement, and then by checking small cases, was able to prove that in fact $p_n > n \log n$ for every n. The paper of Rosser and Schoenfeld (1962) is filled with highly useful and numerically explicit inequalities of this kind.

Let us imagine for a moment that the Riemann hypothesis had been proved. Mathematics is never "used up," there is always that next problem around the bend. Even if we know that all of zeta's nontrivial zeros lie on the line $\operatorname{Im} s = \frac{1}{2}$, we can still ask how they are distributed on this line. We have a fairly concise understanding of how many zeros there should be up to a given height T. In fact, as already found by Riemann, this count is about $(1/2\pi)T \log T$. Thus, on average, the zeros would tend to get closer and closer with about $(1/2\pi)\log T$ of them in a unit interval near height T.

This tells us the average distance, or spacing, between one zeta zero and the next, but there is much more that one can ask about how these spacings are distributed. In order to discuss this question, it is very convenient to "normalize" the spacings, so that the average (normalized) gap between consecutive zeros is 1. By Riemann's result, together with our assumption of the Riemann hypothesis, this can be done if we multiply a gap near T by $(1/2\pi)\log T$, or, equivalently, if for each zero ρ we replace its imaginary part $t = \operatorname{Im} \rho$ by $(1/2\pi)t \log t$. In this way we arrive at a sequence $\delta_1, \delta_2, \ldots$ of normalized gaps between consecutive zeros, which on average are about 1.

Checking numerically, we see that some δ_n are large, with others close to 0; it is just the average that is 1. Mathematics is well equipped to study random phenomena, and we have names for various probability distributions, such as Poisson, Gaussian, etc. Is this what is happening here? These zeta zeros are not random at all, but perhaps thinking in terms of randomness has promise.

In the early twentieth century, Hilbert and Pólya suggested that the zeros of the zeta function might correspond to the eigenvalues of some operator. Now this is provocative! But what operator? Some 50 years later in a now famous conversation between Dyson and Montgomery at the Institute for Advanced Study, it was conjectured that the nontrivial zeros behave like the eigenvalues of a random matrix from the so-called *Gaussian unitary ensemble*. This conjecture, now known as the GUE conjecture, can be numerically tested in various ways. Odlyzko has done this, and found persuasive evidence for the conjecture: the higher the batches of zeros one looks at, the more closely their distribution corresponds to what the GUE conjecture predicts.

For example, take the 1 041 417 089 numbers δ_n with n starting at $10^{23} + 17\,368\,588\,794$. (The imaginary parts of these zeros are around 1.3×10^{22}.) For each interval $(j/100, (j+1)/100]$ we can compute the proportion of these normalized gaps that lie in this interval, and plot it. If we were dealing with eigenvalues from a random matrix from the GUE, we would expect these statistics to converge to a certain distribution known as the Gaudin distribution (for which there is no closed formula, but which is easily computable). Odlyzko has kindly supplied me with the graph in Figure 1, which plots the Gaudin distribution against the data just described (but leaves out every second data point to avoid clutter). Like pearls on a necklace! The fit is absolutely remarkable.

The vital interplay of thought experiments and numerical computation has taken us to what we feel is a deeper understanding of the zeta function. But

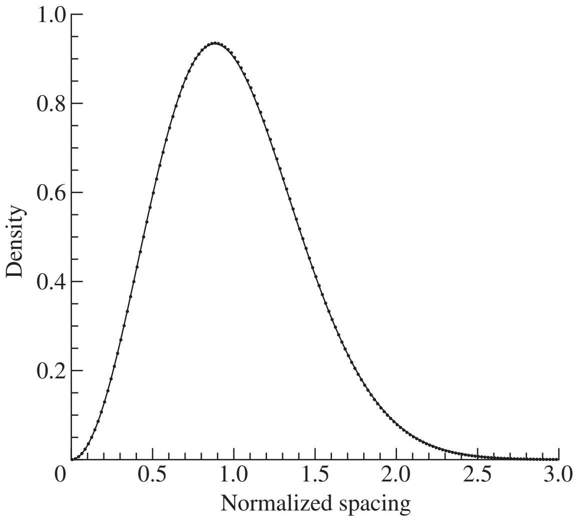

Figure 1 Nearest neighbor spacing.

where do we go next? The GUE conjecture suggests a connection to random matrix theory, and pursuing further connections seems promising to many. It may be that random matrix theory will allow us only to formulate great conjectures about the zeta function, and will not lead to great theorems. But then again, who can deny the power of a glimpse at the truth? We await the next chapter in this development.

5 Diophantine Equations and the *abc* Conjecture

Let us move now from the Riemann hypothesis to Fermat's last theorem. Until the last decade it too was one of the most famous unsolved problems in mathematics, once even having a mention on a *Star Trek* episode. The assertion is that the equation $x^n + y^n = z^n$ has no solutions in positive integers x, y, z, n, where $n \geqslant 3$. This conjecture had remained unproved for three-and-a-half centuries until Andrew Wiles published a proof in 1995. In addition, perhaps more important than the solution of this particular Diophantine equation (that is, an equation where the unknowns are restricted to the integers), the centuries-long quest for a proof helped establish the field of algebraic number theory. And the proof itself established a long-sought and wonderful connection between modular forms and elliptic curves.

But do you know why Fermat's last theorem is true? That is, just in case you are not an expert on all of the intricacies of the proof, are you surprised that there are in fact no solutions? In fact, there is a fairly simple heuristic argument that supports the assertion. First note that the case $n = 3$, namely $x^3 + y^3 = z^3$, can be handled by elementary methods, and this in fact had already been done by Euler. So, let us focus on the cases when $n \geqslant 4$.[1] Let $\mathcal{S}_n$ be the set of positive nth powers of integers. How likely is it that the sum of two members of $\mathcal{S}_n$ is itself a member of $\mathcal{S}_n$? Well, not at all likely, since Wiles has proved that this never occurs! But recall that we are trying to think naively.

Let us try to mimic our situation by replacing the set $\mathcal{S}_n$ with a random set. In fact, we will throw all of the powers together into one set. Following an idea of Erdős and Ulam (1971) we create a set $\mathcal{R}$ by a random process: each integer m is considered independently, and the chance it gets thrown into $\mathcal{R}$ is proportional to $m^{-3/4}$. This process would typically give us about $x^{1/4}$ numbers in $\mathcal{R}$ in the interval $[1, x]$, or at least this would be the order of magnitude. Now the total number of fourth and higher powers between 1 and x is also about $x^{1/4}$, so we can take our random set $\mathcal{R}$ as modelling the situation for these powers, namely the union of all sets $\mathcal{S}_n$ for $n \geqslant 4$. We ask how likely it is to have $a + b = c$ where a, b, c all come from $\mathcal{R}$.

The probability that a number m may be represented as $a + b$ where $0 < a < b < m$ and $a, b \in \mathcal{R}$ is proportional to $\sum_{0<a<m/2} a^{-3/4}(m-a)^{-3/4}$, since for each a less than m the probability that a and $m-a$ both lie in $\mathcal{R}$ is $a^{-3/4}(m-a)^{-3/4}$. Actually, there is a minor caveat when m is even, since then $a = m - a$ when $a = \frac{1}{2}m$: to cover this, we add the single term $(\frac{1}{2}m)^{-3/4}$ to the above sum. Replacing each $m - a$ in the sum with $\frac{1}{2}m$, we get a larger sum that is easy to estimate and turns out to be proportional to $m^{-1/2}$. That is, the chance that a random number m is a sum of two members of $\mathcal{R}$ is at most a certain quantity that is proportional to $m^{-1/2}$. Now the events that would have to occur for m to be given as such a sum involve numbers smaller than m, so the event that m itself is in $\mathcal{R}$ is independent of these. Therefore, the probability that m is not only the sum of two members of

[1] Actually, Fermat himself had a simple proof in the case $n = 4$, but we ignore this.

$\mathcal{R}$, but also itself a member of $\mathcal{R}$, is at most a quantity proportional to $m^{-1/2}m^{-3/4} = m^{-5/4}$. So now we can count how many times we should expect a sum of two members of $\mathcal{R}$ to itself be a member of $\mathcal{R}$. This is at most a constant times $\sum_m m^{-5/4}$. But this sum is convergent, so we expect only finitely many examples. Further, since the tail of a convergent series is tiny, we do not expect any large examples.

Thus, this argument suggests that there are at most finitely many positive integer solutions to

$$x^u + y^v = z^w, \tag{2}$$

where the exponents u, v, w are at least 4. Since Fermat's last theorem is the special case when $u = v = w$, we would have at most finitely many counterexamples to that as well.

This seems tidy enough, but now we get a surprise! There are actually *infinitely many solutions* to (2) in positive integers with u, v, w all at least 4. For example, note that $17^4 + 34^4 = 17^5$. This is the case $a = 1$, $b = 2$, $u = 4$ of a more general identity: if a, b are positive integers, and $c = a^u + b^u$, we have $(ac)^u + (bc)^u = c^{u+1}$. Another way to get infinitely many examples is to build on the possible existence of just one example. If x, y, z, u, v, w are positive integers satisfying (2), then with the same exponents, we may replace x, y, z with $a^{vw}x$, $a^{uw}y$, $a^{uv}z$ for any integer a, and so get infinitely many solutions.

The point is that events of the kind that we are considering—that a given integer is a power—are not quite independent. For instance, if A and B are both uth powers, then so is AB, and this idea is exploited in the infinite families just mentioned.

So how do we neatly bar these trivialities and come to the rescue of our heuristic argument? One simple way to do this is to insist that the numbers x, y, z in (2) be relatively prime. This gives no restriction whatsoever in the Fermat case of equal exponents, since a solution to $x^n + y^n = z^n$ with d the greatest common divisor of x, y, z leads to the coprime solution $(x/d)^n + (y/d)^n = (z/d)^n$.

Concerning Fermat's last theorem, one might ask how far it had actually been verified before the final proof by Wiles. The paper by Buhler et al. (1993) reports a verification for all exponents n up to 4 000 000. This type of calculation, which is far from trivial, has its roots in nineteenth century work of Kummer and early twentieth century work of Vandiver. In fact, Buhler et al. (1993) also verify in the same range a related conjecture of Vandiver dealing with cyclotomic fields, but this conjecture may in fact be false in general.

The probabilistic thinking above, combined with computation of small cases, can carry us deeply into some very provocative conjectures. The above probabilistic argument can easily be extended to suggest that (2) has at most finitely many relatively prime solutions x, y, z over all possible exponent triples u, v, w with $1/u + 1/v + 1/w < 1$. This conjecture has come to be known as the Fermat–Catalan conjecture, since it contains within it essentially Fermat's last theorem and also the Catalan conjecture (recently proved by Mihăilescu) that 8 and 9 are the only consecutive powers.

It is good that we do allow for the possibility that there are *some* solutions, and this is where our main topic of computing comes in. For example, since $1 + 8 = 9$, we have a solution to $x^7 + y^3 = z^2$, where $x = 1$, $y = 2$, and $z = 3$. (The exponent 7 is chosen to insure that the reciprocal sum of the exponents is less than 1. Of course, we could replace 7 by any larger integer, but since in each case the power involved is the number 1, they should all together be considered as just one example.) Here are the known solutions to (2):

$$\begin{aligned}
1^n + 2^3 &= 3^2,\\
2^5 + 7^2 &= 3^4,\\
13^2 + 7^3 &= 2^9,\\
2^7 + 17^3 &= 71^2,\\
3^5 + 11^4 &= 122^2,\\
33^8 + 1\,549\,034^2 &= 15\,613^3,\\
1414^3 + 2\,213\,459^2 &= 65^7,\\
9262^3 + 15\,312\,283^2 &= 113^7,\\
17^7 + 76\,271^3 &= 21\,063\,928^2,\\
43^8 + 96\,222^3 &= 30\,042\,907^2.
\end{aligned}$$

The larger members were found in an exhaustive computer search by Beukers and Zagier. Perhaps this is the complete list of all solutions, or perhaps not—we have no proof.

However, for particular choices of u, v, w, more can be said. Using results from a famous paper of Faltings, Darmon and Granville (1995) have shown that for any fixed choice of u, v, w with reciprocal sum at most 1, there are at most finitely many coprime triples x, y, z solving (2). For a particular choice of exponents, one might try to actually find all of the solutions. If it can be handled at all, this task can involve a delicate interplay between ARITHMETIC GEOMETRY, effective methods in TRANSCENDENTAL NUMBER theory, and good hard computing. In particular, the exponent triple sets $\{2,3,7\}$, $\{2,3,8\}$, $\{2,3,9\}$, and $\{2,4,5\}$ are known to have all their solutions in the above table. See Poonen et al. (forthcoming) for the treatment of the case $\{2,3,7\}$ and links to other work.

The "*abc* conjecture" of Oesterlé and Masser is deceptively simple. It involves positive integer solutions to the equation $a+b=c$, hence the name. To put some meaning into $a+b=c$, we define the "radical" of a nonzero integer n as the product of the primes that divide n, denoting this as $\mathrm{rad}(n)$. So, for example, $\mathrm{rad}(10)=10$, $\mathrm{rad}(72)=6$, and $\mathrm{rad}(65\,536)=2$. In particular, high powers have small radicals in comparison to the number itself, and so do many other numbers. Basically, the *abc* conjecture asserts that if $a+b=c$, then the radical of *abc* cannot be too small. More specifically we have the following.

The *abc* conjecture. *For each $\varepsilon>0$ there are at most finitely many relatively prime positive integer triples a, b, c with $a+b=c$ and $\mathrm{rad}(abc)<c^{1-\varepsilon}$.*

Note that the *abc* conjecture immediately solves the Fermat–Catalan problem. Indeed if u, v, w are positive integers with $1/u+1/v+1/w<1$, then it is easily found that we must have $1/u+1/v+1/w\leqslant 41/42$. Suppose we have a coprime solution to (2). Then $x\leqslant z^{w/u}$ and $y\leqslant z^{w/v}$, so that

$$\mathrm{rad}(x^uy^vz^w)\leqslant xyz\leqslant (z^w)^{41/42}.$$

Thus, the *abc* conjecture with $\varepsilon=1/42$ implies that there are at most finitely many solutions.

The *abc* conjecture has many other marvelous consequences; for a delightful survey, see Granville and Tucker (2002). In fact, the *abc* conjecture and its generalizations can be used to prove so many things that I have joked that it is beginning to resemble a false statement, since a false statement implies everything. But probably the *abc* conjecture is true. Indeed, though a bit harder to see, the Erdős–Ulam probabilistic argument can be modified to provide heuristic evidence for it too.

Basic to this argument is a perfectly rigorous result on the distribution of integers n for which $\mathrm{rad}(n)$ is below some bound. These ideas are worked through in the thesis of van Frankenhuijsen and also the new paper by Stewart and Tenenbaum (forthcoming). Here is a slightly weaker statement than the one suggested by these authors: if $a+b=c$ are relatively prime positive integers and c is sufficiently large, then we have

$$\mathrm{rad}(abc)>c^{1-1/\sqrt{\log c}}. \tag{3}$$

One might wonder how the numerical evidence stacks up against (3). This inequality asserts that if $\mathrm{rad}(abc)=r$, then $\log(c/r)/\sqrt{\log c}<1$. So, let $T(a,b,c)$ denote the test statistic $\log(c/r)/\sqrt{\log c}$. A website maintained by Nitaj (www.math.unicaen.fr/~nitaj/abc.html) contains a wealth of information about the *abc* conjecture. Checking the data, there are quite a few examples with $T(a,b,c)\geqslant 1$, the champion so far being

$$\begin{aligned}
a&=7^2\cdot 41^2\cdot 311^3=2\,477\,678\,547\,239\\
b&=11^{16}\cdot 13^2\cdot 79=613\,474\,843\,408\,551\,921\,511\\
c&=2\cdot 3^3\cdot 5^{23}\cdot 953=613\,474\,845\,886\,230\,468\,750\\
r&=2\cdot 3\cdot 5\cdot 7\cdot 11\cdot 13\cdot 41\cdot 79\cdot 311\cdot 953\\
&=28\,828\,335\,646\,110,
\end{aligned}$$

so that

$$T(a,b,c)=\frac{\log(c/r)}{\sqrt{\log c}}=2.43886\ldots.$$

Is it always true that $T(a,b,c)<2.5$?

One can get carried away with heuristics, forgetting that one is not actually proving a theorem, but making a guess. Heuristics are often based on the idea of randomness, and all bets are off if there is some underlying structure. But how do we know that there is no underlying structure? Consider the case of an "*abcd* conjecture." Here we consider integers a, b, c, d with $a+b+c+d=0$. The condition that the terms be relatively prime now takes on two possible meanings: pairwise relatively prime or no nontrivial common divisor of all four numbers. The first condition seems more in the spirit of the three-term conjecture, but may be a tad too strong in that it disallows using any even numbers. So say we take the four terms with no pair having a common factor greater than 2. Under this condition, our heuristics seem to suggest that for each $\varepsilon>0$, we have

$$\mathrm{rad}(abcd)^{1+\varepsilon}<\max\{|a|,|b|,|c|,|d|\} \tag{4}$$

for at most finitely many cases. But consider the polynomial identity

$$(x+1)^5 = (x-1)^5 + 10(x^2+1)^2 - 8$$

(suggested to me by Granville). If we take x as a multiple of 10, the four terms involved in the identity are pairwise relatively prime except for the last two, which have a common factor of 2. Let $x = 11^k - 1$, which is a multiple of 10. The largest of the four terms is 11^{5k}, and the radical of the product of the four terms is at most

$$110(11^k - 2)((11^k - 1)^2 + 1) < 110 \cdot 11^{3k}.$$

The heuristics are saying that this cannot be, yet here it is right before our eyes!

What is happening is that the polynomial identity is supplying an underlying structure. For the four-term *abcd* conjecture, Granville conjectures that for each $\varepsilon > 0$, all counterexamples to (4) come from at most finitely many polynomial families. And the number of polynomial families grows to infinity as ε shrinks to 0.

We have looked here at only a small portion of the field of Diophantine equations, and then we have looked mainly at the dynamic relationship between heuristics and computational searches for small solutions. For much more on the subject of computational Diophantine methods, see Smart (1998).

Heuristic arguments often assume that the objects of study behave as if they were random, and we have visited several cases where it is useful to think this way. Other examples include the twin-prime conjecture (there are infinitely many primes p such that $p+2$ is prime), Goldbach's conjecture (every even number larger than 2 is the sum of two primes), and countless other conjectures in number theory. Often the computational evidence for the probabilistic view is striking, even overwhelming, and we become convinced in the truth of our model. But on the other hand, if it is this pseudo-proof that is all we have to go on, we may still be very far from the truth. Nevertheless, the interplay of computations and heuristic thinking form an indispensable part of our arsenal, and mathematics is the richer for it.

Remarks and Acknowledgments

I would like to recommend to the reader the book by Cohen (1993) for a discussion of computational algebraic number theory, a subject that is neglected in this article. I am grateful to the following people who generously shared their expertise: X. Gourdon, A. Granville, A. Odlyzko, E. Schaefer, K. Soundararajan, C. Stewart, R. Tijdeman, and M. van Frankenhuijsen. I am also thankful to A. Granville and D. Pomerance for helpful suggestions with the exposition. I was supported in part by NSF grant DMS-0401422.

Further Reading

Agrawal, M., N. Kayal, and N. Saxena. 2004. PRIMES is in P. *Annals of Mathematics* 160:781–793.

Buhler, J., R. Crandall, R. Ernvall, and T. Metsänkylä. 1993. Irregular primes and cyclotomic invariants to four million. *Mathematics of Computation* 61:151–153.

Cohen, H. 1993. *A Course in Computational Algebraic Number Theory.* Graduate Texts in Mathematics, Volume 138. Springer.

Crandall, R. and C. Pomerance. 2005. *Prime Numbers: A Computational Perspective*, 2nd edn. Springer.

Darmon, H. and A. Granville. 1995. On the equations $z^m = F(x,y)$ and $Ax^p + By^q = Cz^r$. *Bulletin of the London Mathematical Society* 27:513–543.

Erdős, P. and S. Ulam. 1971. Some probabilistic remarks on Fermat's last theorem. *Rocky Mountain Journal of Mathematics* 1:613–616.

Granville, A. and T. J. Tucker. 2002. It's as easy as *abc*. *Notices of the American Mathematical Society* 49:1224–1231.

Lenstra Jr, H. W., and C. Pomerance. Forthcoming. Primality testing with Gaussian periods. (Available at www.math.dartmouth.edu/~carlp.)

Odlyzko, A. M., and A. Schönhage. 1988. Fast algorithms for multiple evaluations of the Riemann zeta function. *Transaction of the American Mathematical Society* 309:797–809.

Poonen, B., E. Schaefer, and M. Stoll. Forthcoming. Twists of $X(7)$ and primitive solutions to $x^2 + y^3 = z^7$. *Duke Mathematics Journal.* (In the press.)

Rosser, J. B., and L. Schoenfeld. 1962. Approximate formulas for some functions of prime numbers. *Illinois Journal of Mathematics* 6:64–94.

Smart, N. 1998. *The Algorithmic Resolution of Diophantine Equations.* London Mathematical Society Student Texts, Volume 41. Cambridge University Press.

Stewart, C. L., and G. Tenenbaum. Forthcoming. A refinement of the *abc* conjecture. Preprint.