Big Primes

Factoring Big Integers

Paul Garrett

garrett@math.umn.edu

University of Minnesota

If we want **200-digit prime numbers**, we cannot use the naive method we learned in gradeschool: we would not complete the computation in the lifetime of the universe, even using all the computational power of the whole internet.

Note that these well-known tests for primality confirm that a number is prime exactly by **failing to factor it**.

It turns out that **primality testing is much easier than factoring**.

Factoring big numbers is hard, despite striking (and wacky) modern factorization techniques much better than trial division.

Even more surprising are fast modern **probabilistic primality tests**.

For those who long for absolute certainty, it is possible to construct large primes with accompanying **certificates of primality** indicating how to reprove their primality upon demand.

Facts

The number $\pi(N)$ of primes less than N is

$$\pi(N) \sim \frac{x}{\log x}$$

This is the *Prime Number Theorem* (Hadamard and de la Vallée Poussin, 1896).

Riemann observed (1858) that *if* all the complex zeros of the **zeta function** $\zeta(s) = \sum n^{-s}$ lay on the line $\mathrm{Re}(s) = \frac{1}{2}$ *then* (as refined...)

$$\pi(N) = \frac{x}{\log x} + O(\sqrt{x}\,\log x)$$

The conjecture on the location of the zeros is the **Riemann Hypothesis**.

No result approaching this is known: there is *no* known zero-free region $\mathrm{Re}(s) \geq \sigma$ for $\sigma < 1$.

The Prime Number Theorem uses the non-vanishing of $\zeta(s)$ on $\mathrm{Re}(s) = 1$.

Special Primes

As of January 2000, the largest prime known was the 38^{th} Mersenne prime

$$2^{6972593} - 1$$

Theorem *(Lucas-Lehmer)* Define $L_o = 4$, $L_n = L_{n-1}^2 - 2$. Let p be an odd prime. The Mersenne number $2^p - 1$ is **prime** if and only if

$$L_{p-2} = 0 \bmod 2^p - 1$$

Theorem (*Proth*) The Fermat number $F_n = 2^{2^n} + 1$ is **prime** if and only if

$$3^{(F_n-1)/2} = -1 \bmod F_n$$

Hunting for primes

Nevertheless, when developing expectations for hunting for primes, we pretend that primes are distributed as evenly as possible.

Note: it is *not true* that primes are distributed evenly, even under the Riemann Hypothesis.

But if primes *were* evenly distributed, then near x primes would be about $\ln x$ apart.

Thus, in hunting for primes near x expect to examine $\frac{1}{2}\ln x$ candidates:

For $x \sim 10^{20}$ we have $\frac{1}{2}\ln x \sim 23$

For $x \sim 10^{100}$ we have $\frac{1}{2}\ln x \sim 115$

For $x \sim 10^{500}$ we have $\frac{1}{2}\ln x \sim 575$

Failure of trial division:

Trial division attempts to divide a given number N by integers from 2 up through $\sqrt{N}$. Either we find a proper factor of N, or N is prime. (If N has a proper factor ℓ larger than $\sqrt{N}$, then $N/\ell \leq \sqrt{N}$.) The extreme case takes roughly $\sqrt{N}$ steps, or at least $\sqrt{N}/\ln N$.

If $N \sim 10^{200}$ is prime, or if it is the product of two primes each $\sim 10^{100}$, then it will take about 10^{100} trial divisions to discover this. Even if we're clever, it will take more than 10^{98} trial divisions.

If we could do 10^{12} trials per second, and if there were a 10^{12} hosts on the internet, with $< 10^8$ seconds per year, a massively parallel trial division would take ...

$$10^{66} \text{ years}$$

Examples of trial division

What are the practical limitations of trial division? On a 2.5 Gigahertz machine, code in C++ using GMP

1002904102901 has factor 1001401
('instantaneous')

100001220001957 has factor 10000019
(3 seconds)

10000013000000861 has factor 100000007
(27 seconds)

10000001100000000721 has factor 1000000007
(4 minutes)

The Birthday Paradox [sic]

For $n+1$ things chosen (with replacement) from N the probability that they're **all different** is

$$p = (1 - \frac{1}{N})(1 - \frac{2}{N}) \cdots (1 - \frac{n}{N})$$

Then from $\ln(1-x) > -x$ for small x one has

$$\ln p > -\sum_{\ell=1}^{n} \frac{\ell}{N} \sim -\frac{\frac{1}{2}n^2}{N}$$

Thus, to be sure that $p \geq \frac{1}{2}$ it suffices to take n such that

$$n > \sqrt{N} \cdot \sqrt{2 \ln 2} \sim 1.1774 \cdot \sqrt{N}$$

Thus, with 23 people in a room the probability is greater than $\frac{1}{2}$ that two will have the same birthday.

Pollard's rho method (circa 1976)

We'll try to beat the $\sqrt{N}$ steps trial division needs to factor N.

First try: Suppose that $N = p \cdot M$ with p prime and $p < \sqrt{N}$. If we choose somewhat more than $\sqrt{p}$ integers x_i at random, then the probability is $> \frac{1}{2}$ that for some $i \neq j$ we'll have

$$x_i = x_j \bmod p$$

The probability is roughly $\frac{1}{\sqrt{N/p}} \sim 0$ that $x_i = x_j \bmod N$, so most likely for **some** pair

$$\gcd(x_i - x_j, N) = \text{proper factor of } N$$

But we might have to compare

$$\sqrt{p} \cdot \sqrt{p} = p \sim \sqrt{N}$$

pairs, no better than trial division.

(In any case, we compute *gcd*'s quickly by the **Euclidean algorithm**.)

Second try at Pollard's rho

Since we would have had trouble making a large number of truly random choices anyway, let's stipulate that we choose the x_i's in a more structure way, in a sort of **random walk** in $\mathbf{Z}/N$. Let $f : \mathbf{Z}/N \to \mathbf{Z}/N$ be a deterministic **'random' function**, fix x_o, and define

$$x_{i+1} = f(x_i)$$

Since f is deterministic

$$f(x_i) = x_j \quad \Longrightarrow$$

$$f(x_{i+1}) = f(f(x_i)) = f(x_j) = x_{j+1}$$

So if the walk enters a **cycle** it stays there.
We use **Floyd's cycle-detection trick**:

Floyd's cycle-detection trick:

Fix x_o, define $y_o = x_o$, and define

$$x_{i+1} = f(x_i) \qquad y_{i+1} = f(f(y_i))$$

so the y_i's take the same walk but twice as fast.

Once the cycle is entered, the y's walk one unit faster than the x's, so in fewer additional steps than the cycle length, $x_j = y_j \bmod p$.

In summary: the initial walk plus cycle takes $\sqrt{p} \leq N^{1/4}$ steps, and another $\sqrt{p}$ for the y's to catch the x's modulo p, so

$$2\sqrt{p} \leq 2N^{1/4} \quad \textbf{steps}$$

to find the factor p of N.

Examples of Pollard's rho factorization

Take $x_o = 2$ and $f(x) = x^2 + 2 \bmod N$ (*this is random...?*). In less than 10 seconds total,

2661 steps to find factor

10000103 of 100001220001957

14073 steps to find factor

100000007 of 10000013000000861

9630 steps to find factor

1000000103 of 10000001100000000721

(Even larger...) 129665 steps for factor

10000000019 of 100000001220000001957

162944 steps for factor

100000000103 of 1000000001060000000309

Yet larger...

89074 steps for

$$1000000000039 \text{ of } 100000000016000000004719$$

12 seconds, 584003 steps for

$$10000000000037 \text{ of } 1000000000016600000004773$$

2 minutes, 5751662 steps for

$$100000000000031 \text{ of } 10000000000016400000004123$$

Modern factorization methods

Since the 1970's, better methods have been found (but not polynomial-time):

quadratic sieve: the most elementary of modern factorization methods, and still very good by comparison to other methods. Descended from Dixon's algorithm.

elliptic curve sieve: to factor n, this replaces the group $\mathbf{Z}/n^\times$ with an elliptic curve E defined over $\mathbf{Z}/n$. In effect, the difference between $\mathbf{Z}/n^\times$ and $\mathbf{Z}/n - \{0\}$ is what indicates that n is composite, and an analogous discrepancy in the case of elliptic curves can be similarly exploited.

number field sieve: Descended from several sources, including Adleman, Pomerance, and Rumely (1983), which made novel use of exponential sums (hence, of irrational algebraic numbers)

Bargain-basement Primality Test: Fermat pseudoprimes

Fermat's Little Theorem asserts that for p prime, $b^p = b \bmod p$.

Proven by induction on b, using

$$(b+1)^p = b^p + \binom{p}{1} b^{p-1} + \ldots + \binom{p}{p-1} b + 1$$

$$= b^p + 1 \bmod p$$

The binomial coefficients are *integers*, and on the other hand, they are divisible by p, since

$$\binom{p}{i} = \frac{p!}{i!\,(p-i)!}$$

and the denominator has no factor of p. (*Unique Factorization...*)

Thus, if n is an integer and $b^n \neq b \bmod n$ for some b, then n is **composite**.

The *converse* is false, but not *very* false...

The only *non-prime* $n < 5000$ with
$2^n = 2 \bmod n$ are 341 561 645 1105 1387
1729 1905 2047 2465 2701 2821 3277 4033
4369 4371 4681

Requiring also $3^n = 3 \bmod n$ leaves 561
1105 1729 2465 2701 2821

Requiring also $5^n = 5 \bmod n$ leaves 561
1105 1729 2465 2821

Compared with 669 primes under 5000, this is a *false positive* failure rate of less than 1%.

n is a **Fermat pseudoprime base** b if
$b^n = b \bmod n$.

Terminology

Usage is not consistent.

My usage is that a number that has passed a primality test (Fermat, Miller-Rabin, etc.) is a **pseudoprime**.

Sometimes a *pseudoprime* is meant to be a *non-prime* which has nevertheless passed a primality test such as Fermat. But for large numbers which have passed pseudoprimality tests we may never know *for sure* whether or not they're prime or composite ...

Another usage is to call a number that has passed a test a **probable prime**.

But this is dangerously close to **provable prime**, which is sometimes used to describe primes with accompanying certificates of their primality.

There are only 172 non-prime Fermat pseudoprimes base 2 under 500,000 versus 41,538 primes, a false positive rate of less than 0.41%

There are only 49 non-prime Fermat pseudoprimes base 2 and 3 under 500,000, a false positive rate of less than 0.118%

There are only 32 non-prime Fermat pseudoprimes base 2, 3, 5 under 500,000

There are still 32 non-prime Fermat pseudoprimes base 2, 3, 5, 7, 11, 13, 17 under 500,000

561 1105 1729 2465 2821 6601 8911 10585
15841 29341 41041 46657 52633 62745 63973
75361 101101 115921 126217 162401 172081
188461 252601 278545 294409 314821
334153 340561 399001 410041 449065
488881

Adding more such requirements does not shrink these lists further.

n is a **Carmichael number** if it is a *non-prime* Fermat pseudoprime to *every* base b.

In 1994 Alford, Granville, and Pomerance showed that there are infinitely-many Carmichael numbers.

And it appears that among *large* numbers Carmichael numbers become more common.

Nevertheless, the Fermat test is a very fast way to test for *compositeness*, and is so easy and cheap that it is still the best first approximation to primality.

It is **cheap** because $b^n \bmod n$ can be computed in $\sim \log n$ steps, not n...

Fast modular exponentiation

To compute $b^n \bmod n$, with $n \sim 10^{100}$ or larger, do **not** multiply 10^{100} times.

Rather, note that **repeated squaring** reduces the number of operations:

$$b^{69} = b^{2^6+2^2+2^0} = (((((b^2)^2)^2)^2)^2)^2 \cdot (b^2)^2 \cdot b$$

To compute $x^e \bmod n$

initialize $(X, E, Y) = (x, e, 1)$
while $E > 0$
    if E is even
        replace X by $X^2 \bmod n$
        replace E by $E/2$
    elsif E is odd
        replace Y by $X \cdot Y \bmod n$
        replace E by $E - 1$

The final value of Y is $x^e \bmod n$.

Better primality test: Miller-Rabin (1978)

If $n = r \cdot s$ is composite (with $\gcd(r, s) = 1$) then by Sun-Ze's theorem there are at least 4 solutions to

$$x^2 = 1 \bmod n$$

namely the 4 choices of sign in

$$x = \pm 1 \bmod r \quad x = \pm 1 \bmod s$$

Thus, if we find $b \neq \pm 1 \bmod n$ such that $b^2 = 1 \bmod n$, n is definitely *not* composite.

Roughly, the **Miller-Rabin test** looks for such extra square roots of 1 modulo n (details below).

Theorem: (Miller-Rabin) For composite n, at least $3/4$ of b in the range $1 < b < n$ will detect the compositeness (via the Miller-Rabin test)

Pseudo-corollary If n passes the Miller-Rabin test with k random bases b, then (*exercise: explain the fallacy*)

$$\text{probability}(n \text{ is prime}) \geq 1 - \left(\frac{1}{4}\right)^k$$

Miller-Rabin test base b:
factor $n - 1 = 2^s \cdot m$ with m odd
replace b by $b^m \bmod n$
    if $b = \pm 1 \bmod n$ **stop:** n is 3/4 prime
    else continue
set $r = 0$
while $r < s$
    replace b by $b^2 \bmod n$
    if $b = -1 \bmod n$ **stop:** n is 3/4 prime
    elsif $b = +1 \bmod n$ **stop:** n is composite
    else continue
    replace r by $r + 1$
if we fall out of the loop, n is composite.

If n passes this test it is a
strong pseudoprime base b.

Failure rate of Miller-Rabin?

The fraction of b's which detect compositeness is apparently much greater than 3/4. For $n = 21311$ the detection rate is 0.9976. For 64777 the detection rate is 0.99972. For 1112927 the detection rate is 0.9999973

For $n < 50,000$ there are only 9 non-prime strong pseudoprimes base 2, namely 2047 3277 4033 4681 8321 15841 29341 42799 49141

For $n < 500,000$ there are only 33 non-prime strong pseudoprimes base 2.

For $n < 500,000$ there are *no* non-prime strong pseudoprimes base 2 and 3

For $100,000,000 < n < 101,000,000$ there are 3 strong pseudoprimes base 2 whose compositeness is detected base 3, namely 100463443 100618933 100943201

Some big strong pseudoprimes
Primality testing Fermat base 2, Miller-Rabin base 2, 3, 5, to find next prime after...

('instantaneous')
First prime after 10^{21} is $10^{21} + 117$

('instantaneous')
First prime after 10^{50} is $10^{50} + 151$

('hint of time taken')
First prime after 10^{100} is $10^{100} + 267$

(3 seconds)
First prime after 10^{200} is $10^{200} + 357$

(8 seconds)
First prime after 10^{300} is $10^{300} + 331$

(97 seconds)
First prime after 10^{1000} is $10^{1000} + 453$

Primality Certificates

With origins in work of Eduard Lucas in 1876 and 1891, a very simple form of the **Pocklington-Lehmer theorem** asserts that N is prime if we have

a factorization $N - 1 = p \cdot U$
where p is prime
where $p > \sqrt{N}$
b with $b^{N-1} = 1 \bmod N$
but $\gcd(b^U - 1, N) = 1$

The factorization $N - 1 = p \cdot U$ and the b is the simplest instance of a **certificate of primality** for N.

This requires recursive certification of the prime p.

(The Lucas-Lehmer and Proth criteria are cousins of this idea.)

Lemma *(Fermat, Euler)* For a positive integer N, let b be such that $b^{N-1} = 1 \bmod N$ but $\gcd(b^{(N-1)/p} - 1, N) = 1$. Then a prime divisor q of N satisfies $q = 1 \bmod p$

Proof of lemma: As $b \cdot b^{N-2} = 1 \bmod N$ it must be that b is prime to N, so b is prime to q. Let t be the order of b in $\mathbf{Z}/q^{\times}$. By Fermat's Little Theorem $b^{q-1} = 1 \bmod q$, so $t|q-1$. But the *gcd* condition implies that

$$b^{(N-1)/p} \neq 1 \bmod q$$

Thus, t does not divide $(N-1)/p$. Yet, $t|N-1$. Thus, $p|t$. From $t|q-1$ and $p|t$ we get $p|q-1$, or $q = 1 \bmod p$. ///

Proof of theorem

(Note that if N is prime then $\mathbf{Z}/N$ has a primitive root b which fulfills the condition of the theorem.)

If the conditions of the theorem are met, then all divisors of N are 1 modulo p. If N were not prime, it would have a prime divisor q in the range $1 < q \leq \sqrt{N}$. But $q = 1 \bmod p$ and $p > \sqrt{N}$ make this impossible. Thus, N is prime. ///

Example

By trial division, $p = 1000003$ is prime.

The first strong pseudoprime above $1000 \cdot p$
of the form $p \cdot U + 1$ is

$$N = 1032003097 = 1032 \cdot p + 1$$

By luck, with $b = 2$

$$2^{N-1} = 1 \bmod N$$

while

$$\gcd(2^{(N-1)/p} - 1, N) = \gcd(2^{1032} - 1, N) = 1$$

Therefore, N is *certified* prime.

Continued Example

Let p be the certified prime 1032003097.

The first strong pseudoprime above $10^9 \cdot p$ of the form $p \cdot U + 1$ is
$N = 10320032476724521 63$ which is

$$N = p \cdot (10^9 + 146) + 1$$

By luck, with $b = 2$

$$2^{N-1} = 1 \bmod N$$

while

$$\gcd(2^{(N-1)/p} - 1, N) = 1$$

Therefore, N is *certified* prime.

Continued

Let p be the certified prime
10320032476724 52163

The first strong pseudoprime N above $10^{17} \cdot p$ of the form $p \cdot U + 1$ is

$$p \cdot (10^{17} + 24) + 1$$

$$= 1032003247672452410680779441388519 13$$

By luck, with $b = 2$

$$2^{N-1} = 1 \bmod N$$

while

$$\gcd(2^{(N-1)/p} - 1, N) = 1$$

Therefore, N is *certified* prime.

Continued

Let p be the certified prime
1032003247672452410680779441388519 13

The first strong pseudoprime N above $10^{34} \cdot p$ of the form $p \cdot U + 1$ is

$$p \cdot (10^{34} + 224) + 1$$

$$= \begin{array}{l} 1032003247672452410680779441388 \\ 5422246872747862933999249459487 \\ 102828513 \end{array}$$

By luck, with $b = 2$

$$2^{N-1} = 1 \bmod N$$

while

$$\gcd(2^{(N-1)/p} - 1, N) = 1$$

Therefore, N is *certified* prime.

Continued

Let p be the certified prime

$$\begin{aligned}&10320032476724524106807794413885422\\&46872747862933999249459487102828513\end{aligned}$$

The first strong pseudoprime N above $10^{60} \cdot p$ of the form $p \cdot U + 1$ is (computing for about 5 seconds)

$$\begin{aligned}p \cdot (10^{60} + 1362) + 1 &= 10320032476724524106807794413\\&88542246872747862933999249460\\&89269125184288018334722159917\\&11945402406825893161069777638\\&21434052434707\end{aligned}$$

By luck, $b = 2$ works again and N is *certified* prime.

Continued

Let p be the certified prime

$$10320032476724524106807794413$$
$$88542246872747862933999249460$$
$$89269125184288018334722159917$$
$$11945402406825893161069777638$$
$$21434052434707$$

The first strong pseudoprime N above $10^{120} \cdot p$ of the form $p \cdot U + 1$ is (computing a few seconds)

$$p \cdot (10^{120} + 796) + 1 =$$

$$10320032476724524106807794413885422$$
$$46872747862933999249460892691251$$
$$42880183347221599171194540240682$$
$$93161069777638222555270198542721$$
$$18901900435345279628510707298895463$$
$$40257087058223646693262594438839294$$
$$02708540315833410956211543000018615$$
$$05738026773$$

$b = 2$ works again and N is *certified* prime.

Fast deterministic test for primality

In 2002, Agarwal, Kayal, and Saxena announced a fast (i.e., polynomial time) **deterministic** algorithm for primality testing.

Their algorithm has been checked by a number of experts, including Pomerance.

Still, their algorithm is much slower than the probabilistic Miller-Rabin test.

And there has been recent progress in fast deterministic construction of *random certifiable* primes by Peter Smith, improving Maurer's probabilistic method, and approaching the speed of Miller-Rabin.