Structure and randomness in the prime numbers

A small selection of results in number theory

Science colloquium

January 17, 2007

Terence Tao (UCLA)

Prime numbers

A prime number is a natural number larger than 1 which cannot be expressed as the product of two smaller natural numbers.

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37,$$

$$41, 43, 47, 53, 59, 61, 67, 71, 73, 79, \ldots$$

They are the "atomic elements" of natural number multiplication:

> **Fundamental theorem of arithmetic:** (Euclid, $\approx$ 300BCE) Every natural number larger than 1 can be expressed as a product of one or more primes. This product is unique up to rearrangement.

For instance, 50 can be expressed as $2 \times 5 \times 5$ (or $5 \times 5 \times 2$, etc.).

[It is because of this theorem that we do not consider 1 to be prime.]

Prime numbers were first studied rigorously by the ancient Greeks. One of the first theorems they proved was

Euclid's theorem (≈ 300 BCE) There are infinitely many prime numbers.

Euclid's proof is the classic example of reductio ad absurdum:

- Suppose, for sake of contradiction, that there were only finitely many prime numbers $p_1, p_2, \ldots, p_n$ (e.g. suppose $2, 3, 5$ were the only primes).
- Multiply all the primes together and add (or subtract) 1: $P = p_1 p_2 \ldots p_n \pm 1$. (e.g. $P = 2 \times 3 \times 5 \pm 1 = 29$ or 31.)
- Then P is a natural number larger than 1, but P is not divisible by any of the prime numbers.
- This contradicts the fundamental theorem of arithmetic. Hence there are infinitely many primes.

While there are more direct proofs of Euclid's theorem known today, none are as short or as elegant as this indirect proof.

Euclid's theorem tells us that there are infinitely many primes, but doesn't give us a good recipe for finding them all. The largest explicitly known prime is

$$2^{32,582,657} - 1$$

which is $9,808,358$ digits long and was shown to be prime in 2006 by the GIMPS distributed internet project.

Twin primes

Euclid's proof suggests the following concept. Define a pair of twin primes to be a pair $p, p+2$ of numbers which are both prime. The first few twin primes are

$$(3,5),(5,7),(11,13),(17,19),(29,31),(41,43),\ldots$$

Twin prime conjecture: ($\approx$ 300BCE?)
There are infinitely many pairs of twin primes.

Despite over two millenia of research into the prime numbers, this conjecture is still unsolved! (Euclid's argument suggests that we look for twin primes of the form $p_1 p_2 \dots p_n \pm 1$, but this doesn't always work, e.g. $2 \times 3 \times 5 \times 7 - 1 = 209 = 11 \times 19$ is not prime.)

The largest known pair of twin primes is

$$2,003,663,613 \times 2^{195,000} \pm 1;$$

these twins are $58,711$ digits long and were discovered this Monday (Jan 15, 2007) by Eric Vautier.

The basic difficulty here is that the sequence of primes

$$2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots$$

behaves much more "unpredictably" or "randomly" than, say, the square numbers

$$1, 4, 9, 16, 25, 36, 49, 64, 81, \ldots$$

For instance, we have an exact formula for the n^{th} square number - it is n^2 - but we do not have a (useful) exact formula for the n^{th} prime number p_n!

God may not play dice with the universe, but something strange is going on with the prime numbers. (Paul Erdős, 1913-1996)

Despite not having a good exact formula for the sequence of primes, we do have a fairly good inexact formula:

> **Prime number theorem** (Hadamard, de la Vallée Poussin, 1896) p_n is approximately equal to $n \ln n$. (More precisely: $\frac{p_n}{n \ln n}$ converges to 1 as $n \to \infty$.)

$\ln n$ is the logarithm of n to the natural base $e = 2.71828\ldots$.

This result (first conjectured by Gauss and Legendre in 1798) is one of the landmark achievements of number theory. The proof of this result uses much more advanced mathematics than Euclid's proof, and is quite remarkable:

Very informal sketch of proof:

- Create a "sound wave" (or more precisely, the von Mangoldt function) which is noisy at prime number times, and quiet at other times.

. * *. * . * ... * . * ... * . * ... * .....*

- "Listen" (or take Fourier transforms) to this wave and record the notes that you hear (the zeroes of the Riemann zeta function, or the "music of the primes"). Each such note corresponds to a hidden pattern in the distribution of the primes.

- Show that certain types of notes do not appear in this music. (This is tricky.)
- From this (and tools such as Fourier analysis) one can prove the prime number theorem.

n	p_n	$n \ln n$	Error
10^3	7,919	6,907	-13%
10^6	15,485,863	13,815,510	-10%
10^9	22,801,763,489	20,723,265,836	-9%
10^{12}	29,996,224,275,833	27,631,021,115,928	-8%

The techniques used to prove the prime number theorem can be used to establish several more facts about the primes, e.g.

- All large primes have a last digit of 1, 3, 7, or 9, with a 25% proportion of primes having each of these digits. (Dirichlet, 1837; Siegel-Walfisz, 1963)
  Similarly for other bases than base 10.
- All large odd numbers can be expressed as the sum of three primes. (Vinogradov, 1937)

The **odd Goldbach conjecture** (1742) asserts that in fact all odd numbers n larger than 5 are the sum of three primes.

This is known for $n > 10^{1346}$ (Liu-Wang, 2002) and for $n < 10^{20}$ (Saouter, 1998).

The **even Goldbach conjecture** (Euler, 1742) asserts that all even numbers larger than 2 are the sum of two primes. This remains unsolved.

The prime number theorem asserts that $p_n \approx n \ln n$.

The infamous **Riemann hypothesis** (1859) predicts a more precise formula for p_n, which should be accurate to an error of about $\sqrt{n}$:

$$\int_2^{p_n} \frac{dt}{\ln t} = n + O(\sqrt{n \ln^3 n}).$$

The Clay Mathematics Institute offers a $ 1 million prize for the proof of this hypothesis!

"The music of the primes is a chord"

n	p_n	RH prediction	Error
10^3	7,919	7,773	-1.8%
10^6	15,485,863	15,479,084	$-.04\%$
10^9	22,801,763,489	22,801,627,440	$-.0006\%$
10^{12}	29,996,224,275,833	29,996,219,470,277	$-.00002\%$

Interestingly, the error $O(\sqrt{n \ln^3 n})$ predicted by the Riemann hypothesis is essentially the same type of error one would have expected if the primes were distributed randomly. (The **law of large numbers**.)

Thus the Riemann hypothesis asserts (in some sense) that the primes are pseudorandom - they behave randomly, even though they are actually deterministic.

But there could be some sort of "conspiracy" between members of the sequence to secretly behave in a highly "biased" or "non-random" manner. How does one disprove a conspiracy?

Diffie-Hellman key exchange

Our belief in the pseudorandomness of various operations connected to prime numbers is not purely academic.

One real-world application is Diffie-Hellman key exchange (1976), which is a secure way to allow two strangers (call them Alice and Bob) to share a secret, even when their communication is completely open to eavesdroppers. It, together with closely related algorithms such as RSA, are used routinely in modern internet security protocols.

As an analogy, consider the problem of Alice sending a secret message g by physical mail to Bob, when she suspects that someone is reading both incoming and outgoing mail, and she has no other means of communication with Bob.

Alice can solve this problem as follows.

- Alice writes g on a piece of paper and puts it in a box. She then puts a padlock on that box (keeping the key to herself) and mails the locked box to Bob.
- Bob cannot open the box, of course, but he puts his *own* padlock on the box and mails the doubly locked box back to Alice.
- Alice then unlocks her padlock and mails the locked box back to Bob. Bob then unlocks his own padlock and retrieves the message g.

The (oversimplified) Diffie-Hellman protocol to send a secret number g:

- Alice and Bob agree (over the insecure network) on a large prime p.
- Alice picks a key a, "locks" g by computing $g^a \bmod p$, and sends $g^a \bmod p$ to Bob.
- Bob picks a key b, "double locks" $g^a \bmod p$ by computing $(g^a)^b = g^{ab} \bmod p$, and sends $g^{ab} \bmod p$ back to Alice.
- Alice takes the a^{th} root of g^{ab} to create $g^b \bmod p$, to send back to Bob.
- Bob takes the b^{th} root of $g^b \bmod p$ to recover g.

It is not yet known whether this algorithm is truly secure. (This issue is related to another $ 1 million prize problem: $P \neq NP$.)

However, it was recently shown that the data that an eavesdropper intercepts via this protocol (i.e. $g^a, g^b, g^{ab} \bmod p$) is "uniformly distributed", which means that the most significant digits look like random noise (Bourgain, 2004). This is evidence towards the security of this algorithm.

- Disclaimer 1: The procedure described above is only an oversimplified version of the Diffie-Hellman protocol. The true protocol works slightly differently, generating a "shared secret" g^{ab} for Alice and Bob (and no-one else) only *after* the exchange (in contrast to the secret g used here, which was initially known to Alice but not Bob). This shared secret can then be used as a key to communicate with each other via a standard cipher (such as AES).

- Disclaimer 2: The type of pseudorandomness properties which underlie cryptographic protocols are not the same as the type of pseudorandomness properties which underlie conjectures such as the Riemann hypothesis; thus for instance a solution to the Riemann hypothesis would be a dramatic event in pure mathematics, but would not directly impact cryptographic security.

Sieve theory

The primes are not completely random in their behaviour - they do obey some obvious patterns. For instance, they are all odd (with one exception). They are all adjacent to a multiple of six (with two exceptions). And so forth.

Sieve theory is an efficient way to capture these structures in the primes, and is one of our fundamental tools for understanding the primes.

Sieves study the set of primes in *aggregate*, rather than trying to focus on each prime individually.

They try to "sift out" or "sculpt" the primes by starting with the set of integers and adding or subtracting various components, starting with a few crude and obvious changes, and following up with a many smaller and more subtle changes.

The classic example of a sieve is the Sieve of Eratosthenes ($\approx$ 240BCE), which lets one capture all the primes between $\sqrt{N}$ and N for any given N as follows.

- Start with all the integers between $\sqrt{N}$ and N.
- Throw out (or "sift out") all the multiples of 2.
- Throw out all the multiples of 3.
- ...
- After throwing out all multiples of any prime less than $\sqrt{N}$, the remaining set forms the primes from $\sqrt{N}$ to N.

Modern sieves are more sophisticated, assigning each integer a "score" or "weight" which is upgraded or downgraded depending on what it is a multiple of.

The initial stages of such sieves are easy to understand; it is not hard to compute, for instance, how many numbers, or how many twins, remain after throwing out the multiples of 2 or 3. But the late stages of the sieve are very complicated to deal with.

However, if one terminates the sieve a little earlier (e.g. only throwing out multiples of primes less than $N^{1/4}$ instead of $\sqrt{N}$) then it turns out that it is still possible to keep an accurate count of everything. The catch is that the sieve now captures not only primes, but also almost primes - numbers with very few prime factors. This can be used to give some "near misses" on old conjectures, for instance

Chen's theorem (1966): There exist infinitely many pairs $p, p+2$ where p is a prime and $p+2$ is the product of at most two primes.

Arithmetic progressions of primes

As we mentioned earlier, we are still unable to detect several types of patterns in the primes. However, we have made recent progress on one type of pattern, namely an arithmetic progression $a, a+r, \ldots, a+(k-1)r$.

Green-Tao theorem (2004): The primes contain arbitrarily long arithmetic progressions.

In particular, for any given k, the primes contain infinitely many arithmetic progressions of length k.

This result builds upon a number of existing results; for instance, in 1939, van der Corput showed that the primes contained infinitely many arithmetic progressions $a, a + r, a + 2r$ of length three.

$$2$$

$$2, 3$$

$$3, 5, 7$$

$$5, 11, 17, 23$$

$$5, 11, 17, 23, 29$$

$$7, 37, 67, 97, 127, 157$$

$$7, 157, 307, 457, 607, 757$$

$$\ldots$$

The longest explicitly known arithmetic progression of primes contains twenty-three primes and was discovered by Frind, Jobling, and Underwood in 2004:

$$56,211,383,760,397 + 44,546,738,095,860n;$$

$$n = 0, \ldots, 22$$

Ultra-short, oversimplified sketch of proof

- Using sieve theory one can already show that the almost primes contain long progressions.
- The primes are a subset of the almost primes, but they could be distributed within the almost primes either in a pseudorandom manner or in a structured manner (we don't know which yet).

- However, it is possible to show that in either case, the primes capture a significant fraction of the arithmetic progressions that the almost primes possess.
- (This is a special property of arithmetic progressions, not shared by most other patterns - the property of having lots of these progressions appears to be somewhat "hereditary" and can be passed down to subsets.)

There is still much work to be done. For instance, our theorem shows that the first arithmetic progression of primes of length k has all entries less than

$$2^{2^{2^{2^{2^{2^{100k}}}}}}.$$

(The true size is conjectured to be more like k^k.)

If the Riemann hypothesis is true, we can remove one exponential.