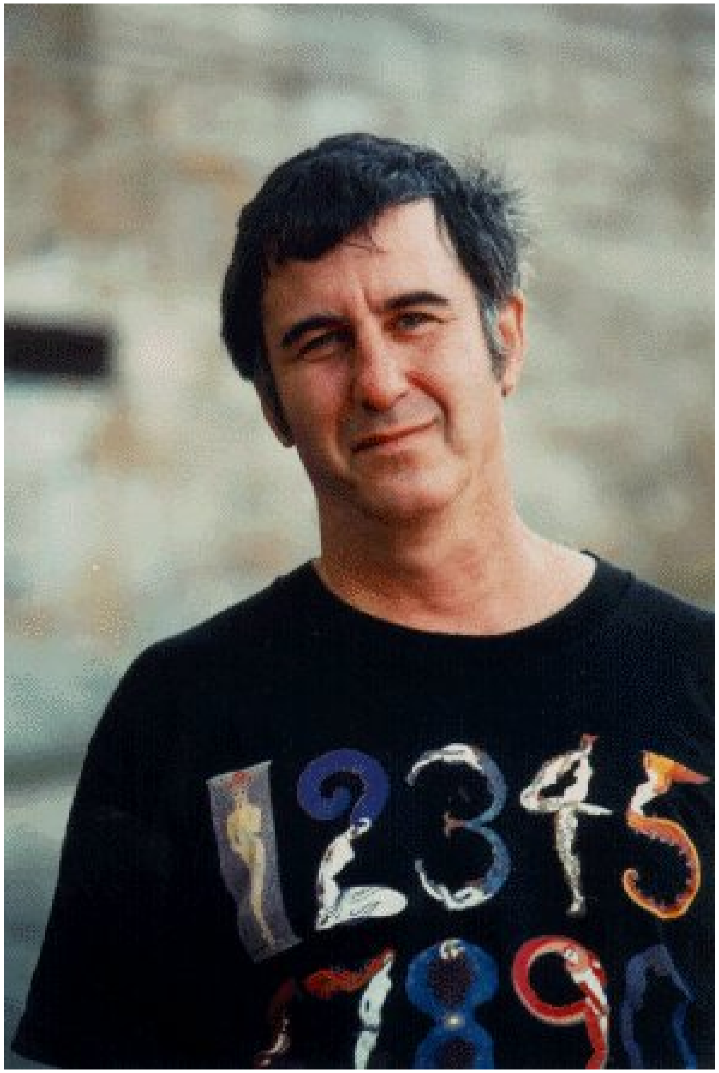
Random Roads:
A celebration of Joel Spencer’s 70th birthday
April 30, 2016

Random number theory

Carl Pomerance, Dartmouth College



Joel Spencer:
closet number theorist

Benkoski & Erdős: A positive integer n is *weird* if the sum of the proper divisors of n exceeds n, yet no sub-sum of these divisors hits n exactly.

They proved that a positive proportion of the natural numbers are weird.

Benkoski & Erdős: A positive integer n is *weird* if the sum of the proper divisors of n exceeds n, yet no sub-sum of these divisors hits n exactly.

They proved that a positive proportion of the natural numbers are weird.

And what is the first weird number?

Benkoski & Erdős: A positive integer n is *weird* if the sum of the proper divisors of n exceeds n, yet no sub-sum of these divisors hits n exactly.

They proved that a positive proportion of the natural numbers are weird.

And what is the first weird number? It is **70**.

In 1770, **Euler** wrote:

"*Mathematicians have tried in vain to discover some order in the sequence of prime numbers, but we have every reason to believe that there are some mysteries which the human mind will never penetrate.*"

from A. Granville, "Harald Cramér and the distribution of prime numbers"

In 1770, **Euler** wrote:

"*Mathematicians have tried in vain to discover some order in the sequence of prime numbers, but we have every reason to believe that there are some mysteries which the human mind will never penetrate.*"

Nevertheless, **Euler** proved in 1737 that the sum of the reciprocals of the primes to x diverges to infinity like $\log\log x$. So, 33 years before his pessimistic statement, he had a glimmer that the mysterious primes might obey some statistical law.

Less than 30 years after **Euler** opined on the mysteries of the primes, **Gauss**, as a teenager, arrived at the conjecture that the number of primes up to x is approximately

$$\int_2^x \frac{\mathrm{d}t}{\log t}.$$

He wrote in 1849 in a letter to Encke:

"*As a boy I considered the problem of how many primes there are up to a given point. From my computations, I determined that the density of primes near x is about $1/\log x$.*"

op. cit.

Here are some notes in Gauss's hand found in the Göttingen library.

Yuri Tschinkel, courtesy of Brian Conrey

Primzahlen

von 1000000 bis 1100000.

	0.	1.	2.	3.	4	5.	6	7.	8	9	
1.	1.										1.
2.		1.				1.		1.	1.		4.
3.		4.	2.	2.	3.	1.	2.	3.	3.	1.	21.
4.	2.	8.	5.	4.	3	6	9.	4.	5.	8.	54.
5	11.	10.	8.	18.	12.	10.	10.	12.	15.	8	114
6	14.	14.	18.	21.	16.	22.	19.	15.	17.	15.	171.
7.	26	17.	23.	23.	24.	24.	17.	22.	20.	21.	217.
8.	19.	19.	21.	7.	14.	15.	20.	17.	15.	17.	164.
9.	11.	13.	9.	13.	14.	14.	12.	13.	11.	16.	126.
10.	8.	6.	8.	5.	9.	5.	5.	9.	7.	9.	71.
11.	6.	6.	4.	6.	3.	1.	3.	1.	4.	5.	39.
12.	1.	1.	2.	1.	1.	1.	2.	2.	1.		12.
13.	1.	1.			1.		1.	1.	1.		6.
14											
15.											
16.											
	752	719	732.	700.	731.	698.	713.	722.	706.	737.	7210.

$\int \frac{dx}{lx} = 7212{,}99$

This conjecture of Gauss may be viewed as saying it is appropriate to study the primes *statistically*.

It led to the **Riemann** Hypothesis (1859). And to the prime number theorem (**Hadamard** & **de la Vallee Poussin** in 1896, **Erdős** & **Selberg** 1949).

More relevant to this talk, this *statistical* view of primes morphed into a *probabilistic* view. In 1923, **Hardy** and **Littlewood** conjectured that the density of twin primes near x is given asymptotically by $c/(\log x)^2$. That is, p and $p+2$ are "independent events" where the constant $c \approx 1.33$ is a fudge factor to take into account the degree to which they're *not* independent.

Actually, in 1871, **Sylvester** came up with a similar heuristic for the number of representations of an even number as a sum of two primes (and so gave a heuristic for Goldbach's conjecture). Hardy and Littlewood returned to this in 1923, but revised Sylvester's constant. The Hardy–Littlewood constant seems to be the "right" one (following both the reasoning for the constant and numerical experiments).

In 1937, **Cramér** gave an explicitly probabilistic heuristic (citing the Borel–Cantelli lemma), that the length of the maximal gap between consecutive primes in $[1, x]$ is $\sim (\log x)^2$. (In 1995, **Granville** revised Cramér's heuristic to take into account certain conspiracies that can deterministically occur among numbers divisible by a small prime, to get that the maximal prime gap is heuristically $\sim c(\log x)^2$, where c is perhaps $2e^{-\gamma} \approx 1.229$.)

Also, the statistical/probabilistic view moved on beyond the primes themselves.

In 1917, **Hardy** and **Ramanujan** proved that the "normal" number of prime factors of an integer near x is $\log\log x$. (This means that for each fixed $\epsilon > 0$, the asymptotic density of the integers n with between $(1-\epsilon)\log\log n$ and $(1+\epsilon)\log\log n$ prime factors is 1.) Though clearly a statistical result, the proof was not.

In 1934, **Turán** gave a new and simple proof of the **Hardy–Ramanujan** theorem, that was based on the second-moment method in probability, but he didn't realize that that is what he had done!

“*When writing Hardy first in 1934 on my proof of the Hardy–Ramanujan theorem, I did not know what Chebyshev’s inequality was and a fortiori on the central limit theorem. Erdős, to my best knowledge, was at that time not aware too. It was Mark Kac who wrote to me a few years later that he discovered when reading my proof in J. LMS that this is basically probability and so was his interest turned to this subject.*”

Letter of Paul Turán to Peter Elliott in 1976, quoted in Elliott’s
“Probabilistic number theory, vol. II”

The distribution of “abundant” numbers (a topic going back to antiquity) was worked out in the 1920s and 1930s by **Schoenberg**, **Davenport** and others, culminating in the **Erdős–Wintner** theorem in 1939.

Also that year, we had the celebrated **Erdős–Kac** theorem on the Gaussian distribution of the number of prime factors of a number.

So was born “probabilistic number theory”, a vital part of analytic number theory.

But what of the “probabilistic method”, where one proves the existence of various strange things by showing that with a suitable probability distribution, there is a positive chance that they exist?

In 1931, **Sidon** wondered how dense a set of positive integers can be if no number has more than 1 intrinsic representation as a sum of two members of the set. (That is, $a + b = n$ is considered as the same representation of n as $b + a$.) And what is the slowest growing function $f(n)$ for a set where every number has at least one representation as a sum of two members, but not more than $f(n)$ representations?

These problems became the subject of much research over the next 30 years, and some of the best theorems were proved via the probabilistic method:

Erdős (1954): One can take $f(n)$ as $c \log n$ for some c.

Erdős (1956): There's a set where every number n has between $c_1 \log n$ and $c_2 \log n$ representations as a sum of two elements.

Still unsolved: Is there a set and a constant $c > 0$ such that every number n has $\sim c \log n$ representations as a sum of two members of the set, as $n \to \infty$?

In Sidon's original problem, he wondered about having at most one intrinsic representation. **Erdős** and **Rényi**, using the probabilistic method in 1960, showed that there is a fairly dense set where every number has a *bounded* number of representations as a sum of two members.

In any event, the probabilistic method felt at home in number theory right from the very beginning!

Let us shift gears to the computer age. If p is an odd prime, the function $x^2 \bmod p$ is $2:1$ for nonzero residues x, so there are exactly $\frac{1}{2}(p-1)$ nonzero squares mod p and exactly $\frac{1}{2}(p-1)$ non-squares mod p. Consider the algorithmic problem of finding one of these non-squares.

For example, for $p = 3$, 2 is a non-square. In fact, 2 works as a non-square for "half" of the primes, namely those that are 3 or 5 mod 8. For the prime 7, 3 is a non-square, and 3 works for the primes that are 5 or 7 mod 12. And so on.

This seems painfully easy! But in fact, we do not have a deterministic polynomial time algorithm that produces a non-square for a given input prime p. (Assuming a generalized form of the Riemann Hypothesis allows us to prove that a certain simple algorithm runs in polynomial time.)

But in practice, no one is concerned with this, because we have a wonderful *random* algorithm that produces a non-square mod p. Namely, choose a random residue r mod p and check to see if it is a square or a non-square mod p (there is a simple polynomial-time check). The probability of success is $\frac{1}{2}$, and so the expected number of trials for success is 2.

This simple example is in fact closely tied to the fundamental problems of factoring polynomials over a finite field, and to primality testing.

For primality testing, we've long known of simple random algorithms that will quickly recognize composite numbers, leading us to strong conjectures that those not revealed as composite are prime. It was only recently that a polynomial time primality test was found (**Agrawal, Kayal, Saxena**), but it's not so computer practical, and the random tests remain as the best choice for practical purposes.

We also use probabilistic reasoning to construct deterministic algorithms.

An example is the *quadratic sieve* factoring algorithm that I found in the early 1980s. The method is almost completely heuristic, assuming numbers produced by a particular quadratic polynomial behave like random numbers of similar size.

(Shhh... No one should tell the large composites about this, they don't know we haven't rigorously proved that the quadratic sieve works, they get factored anyway!)

In fact, this state of affairs is largely true for all practical factoring algorithms, from the Pollard rho method, to the elliptic curve method, and the number field sieve. The elliptic curve method explicitly exploits randomness, but is still a heuristic method. The other algorithms, like the quadratic sieve, are deterministic, but with heuristic, probabilistic analyses.

So far we have considered the distribution of the primes, probabilistic number theory, the probabilistic method in number theory, and the role of randomness in number theoretic algorithms.

Let me conclude with an idiosyncratic problem, one that **Erdős** once proclaimed as perhaps his favorite.

A finite set of integer residue classes is said to form a covering, if the union of the residue classes contains every integer.

Two simple examples: 0 mod 1;
0 mod 2, 1 mod 2

To make this nontrivial, let's rule out the modulus 1, and let's also rule out repeated moduli.

A rule-abiding example:
0 mod 2, 0 mod 3, 1 mod 4, 1 mod 6, 11 mod 12

One can see this works by viewing each as 1 or more classes mod 12. Then 0 mod 2 hits the 6 even classes, 0 mod 3 hits 3 and 9, 1 mod 4 hits 1 and 5, 1 mod 6 hits 7, and 11 mod 12 hits 11.

Erdős conjectured in 1950 that there are coverings with distinct moduli where the least modulus is arbitrarily large.

The current record is held by **Nielsen** (2009) who found a covering with least modulus 40. The moduli only involve the primes to 107, but it has more than 10^{50} of them!

This is nice, but where's the probability?

Let's consider a simple fact. If the moduli used are distinct primes, then they cannot cover, no matter what is chosen as representatives for the residue classes. Why?

Say the moduli are $p_1, p_2, \ldots, p_k$, where these are distinct primes. Being in some residue class modulo one of these primes is an independent event from being in a class for another of them. In fact, the asymptotic density of the integers not covered will be exactly

$$\prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right),$$

which can be arbitrarily close to 0, but cannot be 0.

The exact same argument holds if the moduli $m_1, m_2, \ldots, m_k$ are merely pairwise coprime.

So the **Erdős** covering problem is very much one of extremal cases of *dependent* probabilities!

Some years ago I wondered what the maximal density one can cover using all of the integers in $(x, 2x]$ as moduli. Would it be about

$$\sum_{m\in(x,2x]} \frac{1}{m} \sim \log 2 \qquad \text{or} \qquad \prod_{m\in(x,2x]} \left(1 - \frac{1}{m}\right) \sim \frac{1}{2}$$

or somewhere in between?

Over some years a paper slowly developed of **Filaseta, Ford, Konyagin, P, & Yu** (2007). We proved among many other things that the moduli between x and $2x$ behave asymptotically as if they're independent, that is, one cannot remove more than $\frac{1}{2} + o(1)$ of the integers with them.

Our proof used a lemma that the referee pointed out to us resembles the **Lovász** local lemma. I was quite embarrassed since I first learned of the local lemma some years earlier attending Joel's "Ten Lectures".

I was embarrassed again when at the **Erdős** centennial conference, **Hough** announced his disproof of the **Erdős** covering conjecture! There is a number B such that any covering with distinct moduli must use a modulus at most B. We don't know what B is, but at least we know that $B \geq 40$.

And Hough's proof used our version of the local lemma in a strong way.

There are many more links of number theory to probability, and I haven't even mentioned random number generators. Well, perhaps another time.

Happy Birthday Joel!