Prime Numbers: A Much Needed Gap Is Finally Found

John Friedlander

In May of 2013 the *Annals of Mathematics* accepted a paper [Z], written by Yitang Zhang and showing "bounded gaps for primes," that is, the existence of a positive constant (specifically mentioned was 70 million) with the property that infinitely many pairs of primes differ by less than that constant. Zhang's result created a sensation in the number theory community, but much more broadly as well.

I don't know what it says about the current state of the world or of mathematics or maybe just of me, but I began writing these words by going to Google and typing in "zhang, primes, magazine." Among the first 10 out of more than 74,000 hits, I found references to articles on this topic by magazines with the names *Nautilus, Quanta, Nature, Discover, Business Insider*, and *CNET*. (Within a week of my beginning this, there has appeared a long article [W] in the *The New Yorker*.) I can't begin to guess how many more there have been. I understand that there is also a movie and, I guess, probably television interviews as well. Zhang has since won a number of prizes, including a MacArthur Fellowship, the Ostrowski Prize, the Rolf Schock Prize of the Royal Academy of Sciences (Sweden), and a share of the Cole Prize of the American Mathematical Society. There have also been quite a number of professional papers written about the mathematics and its ensuing developments.

Thus, when I was invited by Steven Krantz to write this article and I requested a few days to think it over, my overriding concern naturally was: "What can I possibly write that is not simply covering well-trodden ground?" This is my excuse for what follows being (I hope) somewhat of a compromise between "folksy gossip" and "bounded gaps for nonspecialists." Perhaps that is what such a *Notices* article should be.

One aspect of this saga—one which presumably should have no place at all—concerns the refereeing of the paper. Both before and after the acceptance of the paper there has been an unseemly amount of attention paid to this, and I admit to being in the process of exacerbating that here. However, the recent article in the *New Yorker* magazine has now driven the final nail in the coffin of confidentiality, the adherence to which has, from the beginning, been what might most charitably be described as tenuous.

As I understand it, the Zhang paper, although received a few days earlier, was first seen by the editor on April 22, 2013. On the morning of April 24, in response to his request of the previous day for a quick opinion, I wrote back, "At first glance this looks serious. I need a few days to think about it and shall write again after that."

A couple of days later I phoned my old friend Henryk Iwaniec to make arrangements regarding a joint project we were going to work on during my trip a few days later to the Institute for Advanced Study in Princeton. "The *Annals* sent me a paper," he mentioned. "I'll bet it's the same one they sent me," I replied. Within a few days, I was at IAS, and meanwhile it seemed the *Annals* had received a raft of quick opinions, all suggesting that there was a nontrivial chance of this being correct, but every one of them deflecting elsewhere the request by the editor to give the paper detailed refereeing.

After Henryk and I, independently and then jointly, had initially refused to do this, we wrote on May 4: "…have been further considering your

John Friedlander is university professor of mathematics at the University of Toronto. His email address is frdlndr@math.toronto.edu.

DOI: http://dx.doi.org/10.1090/noti1257

request that we referee this paper. The paper, if correct, would be of great importance and absolutely deserve to be published in the *Annals*. As the ideas appear to have a very good chance of succeeding, we have decided to accept your invitation to jointly referee it."

That message was well timed, for the next morning we received an email from a friend (initials P.S.) with the words: "…By the way, Katz tells me that he is dealing with a paper at the *Annals* which claims bounded gaps in primes and that all experts (e.g…) are simply passing the buck. That is, you guys are saying it looks like it may be correct (is that so?) but no one is willing to commit to look closely. I haven't seen the paper but if it is serious—it isn't a good sign if everyone suggests each other to referee." I was glad to be in a position wherein we were able to write back: "Before you bawl us out, you might want to talk to Nick again."

I should interrupt this narrative to stress that I am quite sure that almost any of the other "experts" referred to would have, if pressed a little bit harder, agreed to do this job. Henryk and I had the advantage of being on the scene and also of having each other for company.

As many will know, among visitors to the Institute, it can be only a truly monastic individual who is willing to pass up the lunches. So, as a welcome break from the very long hours of checking every line (we had already seen that the basic ideas could not be otherwise dismissed), we went to lunch and sat at the "math table." There were only a few people present there who did not know what we were doing. Our response to the questioning on Monday, "So far, so good." On Tuesday, "So far, so good." On Wednesday, the same. On Thursday, May 8, we were able to say, "It is correct." Two hours later, the report had been sent.

Meanwhile, emails enquiring about the paper's correctness had also been arriving from elsewhere; from D.G. in NY, from D.G. in SJ, from S-T. Y. in Cambridge. One I liked came from Zhang's Princeton namesake: "…I knew him very well even before I came to the US. It would be great if he proved such a theorem." Within a few days, Zhang had his referee report and was giving an invited lecture on his result at Harvard.

I won't write more than a few words about Zhang's history. I can add nothing to the many magazine articles that have been written. A young person raised during a turbulent period of Chinese history, captivated by mathematics but without easy access to library resources, the difficulty in obtaining an education, then even more so an academic position, but continuing to follow the dream—the dream about prime numbers and zeta-functions—then, after many years, finding a stunning success which had evaded the experts. It's a storybook ending to a tale well suited to capturing the public fancy. In addition to showcasing this spectacular theorem, the publicity is of course a very good thing to happen to a discipline that seems not to advertise itself quite as well as do some others and which can be heard to not infrequently complain about the (evident, but perhaps not totally unrelated) inequities in research funding.

In any case, let's pass on to what is always the most beautiful part of the story, the mathematics. The intention here is to make the description brief and accessible to all potential readers. For those interested in a much more extensive introduction to the subject matter yet without forging into the original works, there is now the excellent treatment given in the *Bulletin* article [G] of Andrew Granville.

Everybody knows to whom we credit the origin of the Goldbach conjecture; the name tells you that. Just about as famous is the twin prime conjecture, that there be infinitely many pairs of primes differing by two, such as 3 and 5, 17 and 19, 41 and 43 (yes, I've left some out). Maybe somebody knows the origin of this very old problem, but not I. Although there seems to be no record to substantiate this, it would not be out of the question for Euclid to have speculated about it.

The prime number theorem states that if $\pi(x)$ denotes the number of primes up to x, then, as $x \to \infty$,

$$\pi(x) \sim x/\log x,$$

where the notation means that the quotient of the left side by the right side approaches one as x approaches infinity. This implies that if we are looking at integers, say between x and $2x$, with x large, then the average gap has size about $\log x$. A seemingly modest step beyond this would ask that, for some constants $0 < c < 1 < C$, there be, for arbitrarily large x, pairs of consecutive primes with a gap less than $c \log x$ as well as pairs for which the gap is greater than $C \log x$. A little less modest would be the request that the above hold for *all* c and C satisfying the above inequalities.

For large gaps C, the stronger request was, already in 1931, substantiated by a theorem of Westzynthius. For small gaps, progress came far more slowly, and although the existence of some acceptable $c < 1$ was proven by Erdős in 1940, and despite several highly nontrivial papers further lowering the known verifiable constant, the proof that the result holds for arbitrary positive c came only in the past few years, with the breakthrough paper [GPY] of Goldston, Pintz, and Yıldırım.

Successful approximations to the still unproven twin prime conjecture, as to many problems about prime numbers, start with sieve methods. Although extensively developed over the past century, such methods are, at root, elaborations of the ancient

sieve of Eratosthenes. One takes a finite sequence $\mathcal{A}$ of integers and a set $\mathcal{P}$ of (small) primes and tries to estimate the number of integers in $\mathcal{A}$ not divisible by any prime in $\mathcal{P}$. This is done by exclusion-inclusion and requires us to know, for many given d formed by taking products of subsets of the primes in $\mathcal{P}$, just how many of the integers of $\mathcal{A}$ are divisible by d. For many sequences $\mathcal{A}$ of arithmetic interest, this number has a smooth approximation apart from a small remainder r_d, and just how this "error" accumulates when summed over d determines the degree of success of the argument. In the case of the twin prime problem, one may begin with the set of integers

$$\mathcal{A} = \{p - 2 : p \le x,\ p \text{ prime}\}$$

and cast out multiples of odd integers d, as large as we can, subject to the summed remainder, say $R(D) = \sum_{d \le D} |r_d|$, being acceptable. In this fashion, Renyi succeeded in being the first to prove that, for some fixed r, there are infinitely many primes p such that $p - 2$ has at most r prime factors. The smallest known value of r was eventually lowered to 2 by J-R. Chen, but the so-called parity phenomenon of sieve theory (see [FI2]) prevents one from reaching the goal along these lines.

Note that in the above scenario the number of multiples of a given integer d is just $\pi(x; d, a)$, the number of primes $p \le x$ in the arithmetic progression a modulo d, specifically for $a = 2$. The prime number theorem for arithmetic progressions gives us an asymptotic formula for this quantity, but so far one has succeeded to prove this only for d very small compared to x, and as stated above, it is crucial to have information for many (and hence also for large) d. Fortunately, we need this information for the remainder only on average over d. The Bombieri-Vinogradov theorem tells us that this averaged remainder is small for d almost up to $x^{1/2}$. We have (slightly more than) the bound

$$\sum_{d < x^{\frac{1}{2}-\varepsilon}} \max_{\gcd(a,d)=1} |\pi(x; d, a) - \pi(x)/\varphi(d)| \ll x/(\log x)^A$$

holding for arbitrary fixed positive A and ε. Here, the Euler function $\varphi(d)$ counts the number of "reduced" residue classes modulo d, that is, those relatively prime to the modulus, and the symbol $\ll$ indicates that the left side is bounded by some constant multiple of the right side. The B-V theorem has played a key role in all attacks on the small gaps problem, and indeed Bombieri's work was motivated by an important step along the way [BD] jointly with H. Davenport.

The parity phenomenon tells us that we can't hope to succeed in proving, along these lines, that there are pairs of integers differing by two, both of which are prime; we have to give something up. In the Renyi-Chen theorem we sacrifice one of the integers being prime, but get close to the result in that this integer is an "almost-prime." In the GPY method we insist on requiring both integers to be prime, but we give up the demand that they necessarily have a single prescribed difference, much less that the distance specifically be two. Indeed, such a sacrifice had already been conceded in all earlier attempts at the small gap problem, but GPY introduced ideas that went much further than previously and were the first to get close to the result. They begin with an "admissible" k-tuple of integers $\mathcal{H} = \{h_1, \dots, h_k\}$. By this we mean that for every prime number p, at least one of the residue classes modulo p is missed by every one of the h_i. It is expected, but is presumably very far from our current reach, that every admissible k-tuple of integers $\mathcal{H}$ ought to possess infinitely many translates $n + \mathcal{H}$, *all* of whose integers are primes. (It is easy to see that admissibility is a necessary condition for such a result to hold. To take the simplest example, note that $\{0, 2\}$ is admissible, but $\{0, 1\}$ is not, and indeed every second integer is even.) The GPY approach asks for a more modest conclusion: it seeks to estimate the number of translates $n + \mathcal{H}$ of $\mathcal{H}$ which contain at least two primes. In addition to proving that there exist pairs of primes arbitrarily closer than the average, [GPY] proves, using an essential (to my knowledge not otherwise published) input from Granville and Soundararajan, that any improved exponent beyond $1/2$ in the Bombieri-Vinogradov theorem would imply the bounded gaps theorem.

There had already been in the 1980s in [Fo-I], [BFI] some results that went beyond the exponent $1/2$ in Bombieri–Vinogradov type estimates. However, these results lacked a certain uniformity in the residue class, one that had been present in the earlier B-V theorem. Specifically, as we noted above in connection with the Renyi approach, regardless of which modulus d was being considered, it was always the same initial term, 2 modulo d. For that approach, wherein the residue class is fixed, the results in [Fo-I], [BFI] were satisfactory. However, when one begins, as does GPY, with a more complicated set, a k-tuple of integers $\{h_1, \dots, h_k\}$, then the relevant residue classes modulo d move around as d does and modifications to the arguments are required.

These modifications have now, several years after [GPY], been successfully implemented for three crucial reasons. In the first place, unlike in the original Bombieri-Vinogradov theorem, wherein every reduced class was permitted to enter, the number of residue classes involved in the k-tuple is not very large. In the second place, the movement of these classes is not arbitrary but occurs in a natural

arithmetic fashion; specifically, it is controlled by the Chinese Remainder Theorem. In the third place, Zhang is very talented. In addition to being completely at ease with all of the relevant earlier work (of which there was much), he succeeded in producing a number of crucial innovations of his own.

The full proof, especially when considered from first principles, is a beautiful intermingling of ideas from various parts of the subject: combinatorial, algebraic, analytic, geometric. Deligne's work makes an appearance in bounds for certain exponential sums (along the lines of [FI1], but only after one of Zhang's most significant improvements).

The excitement created by Zhang's result was by no means confined to the popular press. Very shortly after the news was out, a number of small improvements in Zhang's constant 70 million began to appear on the Number Theory arXiv. During that summer, Iwaniec and I had written up our own account of Zhang's work, originally not intended for publication, but later appearing in [FI4]. About the same time, a Polymath project, led by Terence Tao and attracting contributors, some young and some more established, systematically whittled away in far more substantial fashion, employing ever more delicate arguments, algebro-geometric, combinatorial, and computational. For months, on any given day it was dangerous to claim that one knew the latest value of this constant!

Then, in early autumn, things took a sudden turn. At Oberwolfach in October 2013 and on the arXiv two or three weeks later, James Maynard announced a further breakthrough. Simultaneously, Terence Tao found essentially the same results along closely related lines. Maynard's paper has since been published in [M]. In this, he re-proves Zhang's bounded gap result in both simpler and stronger quantitative form. Moreover, he succeeds in producing, for each given m, not just for $m = 2$, the existence of m primes in infinitely many intervals, each having a length bounded by a constant (depending only on m). Precisely, with p_n denoting the nth prime, one has

$$\liminf_n (p_{n+m} - p_n) < Cm^3 e^{4m}$$

for a universal positive constant C. Even after Zhang's result, it seemed amazing to think one would soon see even triples handled.

Maynard's method (as well as that of Tao) begins, as did Zhang's, with the GPY approach but then immediately takes off in a completely different direction. To describe this we need to say a fair bit more about GPY. The starting point for their method rests on consideration of the difference between two sums.

Let $\mathcal{H} = \{h_1, \dots, h_k\}$, as before, be an "admissible" k-tuple of integers. We consider a sum

$$S = \sum_{x<n<2x} \Big(\sum_{1\le j\le k} \chi_p(n + h_j) - \nu \Big) \theta_n = S_1 - S_2$$

say, where χ_p is the characteristic function of prime numbers and θ_n is a certain nonnegative weight whose intelligent choice is key to the argument. Suppose we can prove for some positive ν that we have $S > 0$. Since θ_n is nonnegative, it follows for at least one n that the quantity in parentheses is positive and for that n, the number of primes $n + h_j$ is at least ν (or possibly better, the least integer greater than or equal to ν), and these primes all lie in an interval of length no greater than the diameter of $\mathcal{H}$.

To evaluate the sum S we deal with the two sums S_1, S_2 separately. To have any chance of success in making this difference positive, we are going to have to require the arithmetic function θ_n to have various properties which are impossible to justify in a brief account and which suggest that one look at Selberg sieve weights. One can find a rather full account of the Selberg sieve and the GPY argument, for example, in Chapter 7 of [FI3]. Suffice it to say here that GPY chose weights of the following type:

$$\theta_n = \Big(\sum_d \mu(d) f(d)\Big)^2 .$$

Here μ is the Möbius function, the sum goes over those $d < D$, $d | (n + h_1) \dots (n + h_k)$, and f is of the form $f(d) = F(\log D/d)$ where F is a nice smooth function to be determined. To evaluate each of the sums S_i we open the square and interchange the order of summation, bringing us to the inner sum, now being over n. In the case of S_2 all goes well, but when it comes to S_1, because of the presence of χ_p, the inner sum counts primes in an arithmetic progression, and we need a Bombieri-Vinogradov type result. This limits the choice of D in that we require an acceptable B-V result with moduli up to level D^2 (because of the square in our choice of θ_n). The final problem is to choose F well. GPY actually made a choice not quite optimal but just about as good for the application.

The innovation which allows so much progress in [M] is simply the attachment of a k-dimensional version of the sieve weight used by GPY, one which gives separate individual treatment to each of the elements of the k-tuple. Roughly speaking, his weights look like

$$\theta_n = \Big(\sum_{\mathbf{d}} \Big(\prod_{1\le i\le k} \mu(d_i) \Big) f(d_1, \dots, d_k) \Big)^2 ,$$

where the sum goes over k-tuples $\mathbf{d} = (d_1, \dots, d_k)$ with $d_i | n + h_i$, $1 \le i \le k$, and $\prod_{1\le i\le k} d_i \le D$.

Selberg had many years ago briefly introduced this multidimensional version of his weight but did not make much use of it, and certainly it was not responsible for any of his great advances in the subject. This new weight leads, in the analytic evaluation of the two sums being compared, to some complications vis-à-vis the one-dimensional version in GPY, in particular in the optimal choice of f, but Maynard overcomes these in elegant fashion.

As a result of this new approach, Maynard is able to dispense with all of the most advanced results on the distribution of primes in arithmetic progressions, such as those which form the centerpiece of innovation in Zhang's proof. The Bombieri-Vinogradov theorem is amply sufficient for the qualitative statement of Maynard's results, and I expect, but do not know if anybody has checked, that even the somewhat complicated-looking precursors of the B-V theorem dating back to Renyi might be sufficient. However, for explicit bounds on these gaps, strong statements of Zhang type are important, with the happy consequence being a new Polymath project [P], including Maynard's incorporation into the enterprise and resulting in further quantitative improvements (for twins, we are now down into the hundreds) by a combination of the two approaches. One particularly striking achievement is a result in [P] which is conditional on the assumption of a very strong Generalized Elliott-Halberstam Conjecture (of a type introduced in [BFI]) and concerning the average distribution in arithmetic progressions to very large moduli (d running up to $x^{1-\varepsilon}$) of certain arithmetic functions. As a consequence of this conjecture, the authors of [P] deduce the existence of infinitely many pairs of primes differing by no more than six.

Meanwhile, there have also been further developments in other directions. The Maynard weights offer various possibilities for future research and, especially following the availability of the preprint form of [M], there quickly followed a dozen or more inventive uses of the new ideas to attack different problems, questions on number fields, on polynomials, on primitive roots, on cluster points of normalized prime gaps, on elliptic curves, on large gaps between primes.…One can find more information on these in Section 12 and Appendix B of [G], the electronic version of which became available a few days before I wrote these words. Those few pages, although near the end of Granville's paper, can be read without having gone through the heavier work of the immediately preceding sections and are highly recommended to the readers of this article. Moreover, all of these works can be located through the very extensive list of references in [G].

I enjoyed having had the opportunity to co-organize, with D. Goldston and K. Soundararajan, a November 2014 workshop at the American Institute of Mathematics shortly before its relocation from Palo Alto to San Jose. It was a particular pleasure to hear talks on many of these innovations and especially to see that a very high percentage of them are due to young mathematicians, young mathematicians of both genders.

References

[BD] E. Bombieri and H. Davenport, Small differences between prime numbers, *Proc. Roy. Soc. Ser. A* **293** (1966), 1–18.

[BFI] E. Bombieri, J. B. Friedlander, and H. Iwaniec, Primes in arithmetic progressions to large moduli, *Acta Math.* **156** (1986), 203–251.

[Fo-I] E. Fouvry and H. Iwaniec, Primes in arithmetic progressions, *Acta Arith.* **42** (1983), 197–218.

[FI1] J. B. Friedlander and H. Iwaniec, Incomplete Kloosterman sums and a divisor problem, with an appendix by B. J. Birch and E. Bombieri, *Ann. Math.* **121** (1985), 319–350.

[FI2] ———, What is the parity phenomenon?, *Notices Amer. Math. Soc.* **56** (2009), 817–818.

[FI3] ———, *Opera de Cribro*, Colloq. Pub., 57, Amer. Math. Soc., Providence, RI, 2010.

[FI4] ———, Close encounters among the primes, *Indian J. Pure Appl. Math.* **45** (2014), 633–689.

[GPY] D. A. Goldston, J. Pintz, and C. Y. Yildirim, Primes in tuples I, *Ann. Math.* **170** (2009), 819–862.

[G] A. Granville, Primes in intervals of bounded length, *Bull. Amer. Math. Soc.* **52** (2015) 171–222.

[M] J. Maynard, Small gaps between primes, *Ann. Math.* **181** (2015), 383–413.

[P] D. H. J. Polymath, *Variants of the Selberg sieve, and bounded intervals containing many primes*, preprint.

[W] A. Wilkinson, The pursuit of beauty, Yitang Zhang solves a pure-math mystery, *New Yorker Magazine*, Profiles, February 2, 2015.

[Z] Y. Zhang, Bounded gaps between primes, *Ann. Math.* **179** (2014), 1121–1174.