ARITHMETIC STATISTICS: ELLIPTIC CURVES AND OTHER THINGS

BARRY MAZUR

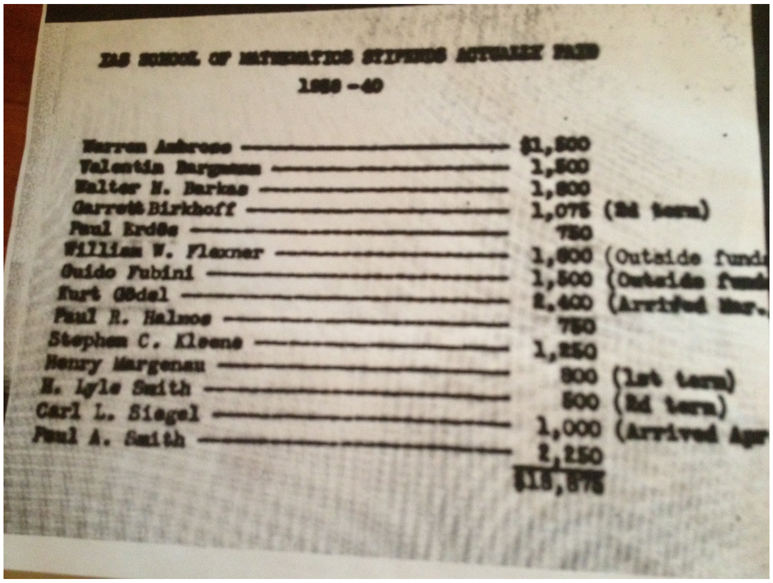
Notes for the Erdös Memorial Lecture, held at Temple University, October 12, 2013

The wonderful thing about giving a lecture called the *Erdös Memorial Lecture* is that there is no need at all to tell anyone who the eponymous mathematician was! And if I dared begin to list his accomplishments we'd be here till the next Erdös Memorial Lecture next year.

The signature *Erdös* on any piece of his mathematics is Ioud and clear. For example, if I flashed on the screen the following challenge:

> **Consecutive early primes: $100/$25000 Prize.** An early prime is one which is less than the arithmetic mean of the prime before and the prime after. Conjecture: There are infinitely many consecutive pairs of early primes. The larger award would be granted for a disproof.

I wouldn't have to say who was the generous benefactor who proposed the problem and offered the reward. There *is* only one person whose style this is! The generosity of that cash reward, by the way, belies his own salary if we can judge by this 1939-40 list of salaries from the Institute for Advanced Study at Princeton.



IAS SCHOOL OF MATHEMATICS STIPENDS ACTUALLY PAID
1939-40

Warren Ambrose	$1,500	
Valentin Bargmann	1,500	
Walter W. Barkas	1,800	
Garrett Birkhoff	1,075	(2d term)
Paul Erdös	750	
William W. Flexner	1,800	(Outside fund
Guido Fubini	1,500	(Outside fund
Kurt Gödel	2,400	(Arrived Mar.
Paul R. Halmos	750	
Stephen C. Kleene	1,250	
Henry Margenau	800	(1st term)
H. Lyle Smith	500	(2d term)
Carl L. Siegel	1,000	(Arrived Apr
Paul A. Smith	2,250	
	$18,875	

I recall being at a conference that Erdös attended (it was either in Warsaw or in Prague, I think, in the very early 60's of the last century). One 'free' afternoon was scheduled during the conference. I remember that almost everyone went off to swim in what was proudly touted to be a very cold lake, except for me–who stayed around to write letters to my girlfriend, and Erdös, who did exactly what he did 24 hours a day: mathematics. For Erdös, mathematics was a 'practice' and he practiced something that might be called a 'mathematics of problems.' He was, in a sense, a sculptor, dealing with a multitude of problems as a sculptor might work with clay, building profound understanding by assembling (and solving) constellations of problems which fit together brilliantly.

Many of his question revolve around the statistics of occurrences of mathematical phenomena in number theory, and this leads to my theme.

For fun—and for focus—this hour, here is an attitude toward Diophantine questions that I don't want to try to defend in any generality. As you will see, it is (hubristically) time-dependent on our present knowledge, and therefore nothing one can really defend. But we will be discussing a few examples illustrating what it leads to, in the way of conjectures, and how far we are toward proving them.

> **A minimalist instinct** Roughly—and statistically—speaking, a (large enough) family of Diophantine equations will have as 'few' rational solutions as it is constrained to have, given what we already know. And everything we don't know behaves 'randomly,' this being taken in some straightforwardly naive sense.

As you see, this dictum is a naive kit for making conjectures–null hypotheses, so to speak. There's little reason to believe a conjecture built from this viewpoint, but it is surprising how often it leads to right answers. In the spirit of Erdös, who made tons of conjectures in his life, let's follow this instinct during this hour.

Note that, foremost, the minimalist instinct leads to statements about families, or 'aggregates' of problems rather than single ones. Why study aggregates?

It is curious how *aggregates* rather than *single instances* creeps into our subject even when we aren't looking for statistical trouble.

Here is an example: in the Erdös spirit, I'll offer a $5 prize for anyone who can manage to provide a proof of the fact that every linear form $aX + b$ with a, b relatively prime represents (for $X \mapsto x \in \mathbf{Z}$) at least *one* prime number and such that the proof doesn't actually show that it represents *infinitely many* primes. I think my $5 is safe, but the point I want to make is that a certain amount of our work is—whether we want it or not—inescapably about "aggregates."

Such aggregates of numbers form essential fodder for number theorists,and there is a real pleasure in just working in the thick of "many numbers," as is vividly expressed in this letter of Gauss to one of his students (the *italics* are mine):

> Even before I had begun my more detailed investigations into higher arithmetic, one of my first projects was to turn my attention to the decreasing frequency of primes, to which end *I counted the primes in several chiliads* and recorded the results on the attached white pages. I soon recognized that behind all of its fluctuations, this frequency is on the average inversely proportional to the logarithm, so that the number of primes below a given bound n is approximately equal to
>
> $$\int dn/\log(n),$$
>
> where the logarithm is understood to be hyperbolic. Later on, when I became acquainted with the list in Vegas tables (1796) going up to 400031, I extended my computation further, confirming that estimate. *In 1811, the appearance of Chernaus cribrum gave me much pleasure and I have frequently (since I lack the patience for a continuous count) spent an idle quarter of an hour to count another chiliad here and there...*

Often, in modern number theory, to actually sample a sufficient quantity of data that might allow you to guess even approximate qualitative behavior of the issue you are studying, you may have to go out to very high numbers. For example, there are basic questions about elliptic curves (e.g., what is the probability that elliptic curves possesses two independent rational points of infinite order?) where if you only look at elliptic curves of conductor $< 10^8$, you might be tempted to make guesses that are not only wrong, but qualitatively wrong.

But let's use our very unsophisticated, and yet unreasonably useful, 'kit' to make some guesses.

1. A version of the ABC Conjecture

A host of conjectures of Erdős have to do with sums of integers that are (relatively high) powers of smaller numbers, or near-powers. The question of whether such a sum can again be a significantly high power has been formulated beautifully by Masser and Oesterlé, and is now called the *ABC Conjecture*. A variant of this problem, formulated by the benefactor of this lecture series, Andrew Beal, has a $1,000,000 prize reward for its solution! (This continues the tradition of Erdös, with extra largess.)

To begin the practice of our minimalist instinct, here, then, is a way of coming to guess a version of the ABC conjecture.

> Let a, b, c be a triple of positive integers. Consider the diophantine equation
>
> $$A + B = C$$
>
> where A, B, and C are positive integers and:
>
> - A is a perfect a-th power,
> - B a perfect b-th power, and
> - C a perfect c-th power.
>
> Let X be a large positive integer, and $N(X)$ be the number of solutions of our diophantine equation with $C \leq X$.
>
> What can we say about the behavior of $N(X)$ as a function of the bound X?

To guess the answer we must:

(1) Deal with any "regularities" that we're aware of; e.g. add the requirement that $GCD(A,B,C)=1$ (which I will ignore in this rough account).

(2) Assume that everything else behaves in an elementary random way. There are:

$\sim\ X^{1/a}$ possible values of A less than X,

$\sim\ X^{1/b}$ possible values of B,

and $\sim\ X^{1/c}$ possible values of C.

So, working with numbers A, B, C less than X we see that we have

$$X^{\frac{1}{a}}\cdot X^{\frac{1}{b}}\cdot X^{\frac{1}{c}}=X^{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$$

shots at achieving a "hit," i.e., such that the value $A+B-C$ is zero.

But $A+B-C$ will range roughly (ignoring multiplicative constants) through X numbers, so the "chance" that we get a hit will be:

$$N(X)\sim\frac{1}{X}\cdot\text{the number of shots}\sim X^{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}-1}.$$

Perhaps this suggests to us that we should guess:

$$X^{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}-1-\epsilon}\quad <<\quad N(X)\quad <<\quad X^{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}-1+\epsilon}\ \ ??$$

But if $\frac{1}{a}+\frac{1}{b}+\frac{1}{c}-1$ is negative we arrive at the ludicrous expectation that $N(X)$ goes to zero as X goes to infinity, suggesting the more civilized guess:

CONJECTURE: If $\frac{1}{a}+\frac{1}{b}+\frac{1}{c}<1$ there are only **finitely many solutions**.

which is in the spirit of the classical ABC conjecture.

2. Elliptic curves

Consider a question raised by Mordell as the title of one of his marvelous papers:

> *What products of two consecutive integers are equal to a product of three consecutive integers?*

The answer to this question, by the way, known to Mordell half a century ago, is that the only such products are $0, 6$, and 210.

The equation whose integral solutions "solves" Mordell's Question is

$$\mathcal{E}: \quad y^2 + y = x^3 - x$$

and this is an affine model, over $\mathbf{Z}$, of an elliptic curve over $\mathbf{Q}$.

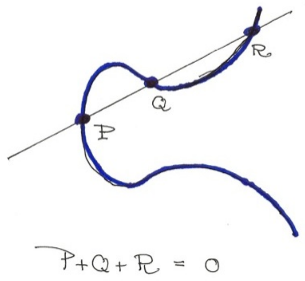
Side-comment: *This elliptic curve knows all other elliptic curves—or at least pairs of elliptic curves related by a* 37*-isogeny—explain!*

Now if you want to know the answer to Mordell's question, you need only study the *integral* solutions of that equation. For such equations (quadratic expressions of the variable y as equal to cubic expressions of x)—-and in contrast to the general problem of integral solutions as posed by Hilbert's Tenth Problem and as solved negatively by Matyasevich—there is an algorithm allowing one to finitely determine all its integral solutions.

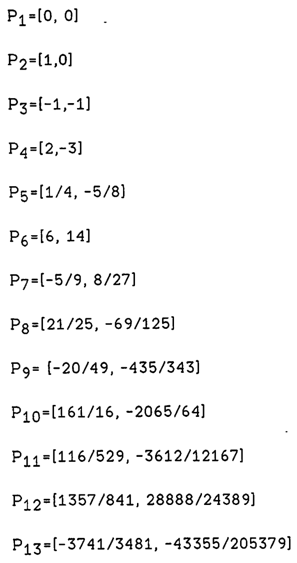
The situation regarding rational solutions of equations defining elliptic curves, however, is quite different. We have a candidate-algorithm, and standard conjectures that say that it will work, but no proof yet. A tiny bit better still: if the algorithm actually does work in any case, and if we persist in computation with that case, we will finally see that it works, but we don't have any a priori estimate for how much time it will take to resolve the question.

If we return to Mordell's equation and ask for its *rational* rather than only integral solutions, we get quite a different, and beautiful, answer: there are infinitely many rational solutions, and all of them are ‘generated’ out of the simplest of its solutions: $(x, y) = (0, 0)$.

The mode of generation was called initially the *chord-and-tangent-process*



which banks on the fact that our curve is a cubic–i.e., of degree 3—and therefore any straight line (in projective space) will intersect it in exactly three points, counting multiplicity, so, for example, the tangent line to the curve at our generating point $P_1 := (0, 0)$ intersects the curve at one other point: $P_2 := (1, 0)$ and so, having now two points we can iterate to get



P_1=[0, 0]

P_2=[1,0]

P_3=[-1,-1]

P_4=[2,-3]

P_5=[1/4, -5/8]

P_6=[6, 14]

P_7=[-5/9, 8/27]

P_8=[21/25, -69/125]

P_9= [-20/49, -435/343]

P_{10}=[161/16, -2065/64]

P_{11}=[116/529, -3612/12167]

P_{12}=[1357/841, 28888/24389]

P_{13}=[-3741/3481, -43355/205379]

and the full answer is that there are infinitely many rational points P_n for n positive negative and 0 (the point at infinity) and, moreover, the chord-and-tangent-process allows us to describe these points on our (projective) plane curve in an elegant recursive way: three points P_a, P_b, and P_c are collinear in the plane if and only their subscripts sum to zero: $a + b + c = 0$.

In general, this type of structure is the key to understanding any elliptic curves E and its rational point over any number field K. Denoting by $E(K)$ its set of rational points, the chord-and-tangent-process endows $E(K)$ with an abelian group structure, and a fundamental theorem (1922) of Mordell (over $\mathbf{Q}$) extended by Andrei Weil over any number field K says that this group $E(K)$ is a *finitely generated* abelian group (called naturally, the **Mordell-Weil group** of E over K) and so is characterized up to isomorphism by its two invariants:

- its *torsion subgroup*, $T(E, K)$,
- and its *rank* $r(E, K)$.

I.e.,

$$E(K) \quad \simeq \quad T(E, K) \quad \bigoplus \quad \mathbf{Z}^{r(E,K)}.$$

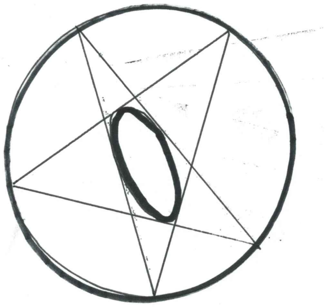
This immediately leads to two mathematical projects that are—as it turns out—surprisingly different.

- Study the behavior of torsion $(E, K) \ \mapsto \ T(E, K)$,
- Study the behavior of rank $(E, K) \ \mapsto \ r(E, K)$,

as functions of varying elliptic curves and number fields.

3. Torsion

Torsion in elliptic curves have, as one of their many neat realizations, periodic arrays in the classical it Poncelet Billiard game where you have a configuration of two conics in the plane (I think of one of them as the "outer conic" comprising the outer profile of the billiard table, encircling the other conic, which we'll call the "inner conic," and which we can think of as an obstruction on the table. The game is to make a shot that bounces multiple times off the rim of the outer conic, but each time it comes back, its path just grazes the inner one, and it makes a finite periodic trajectory this way.



A theorem of mine gave a complete classification of torsion, rational over Q for elliptic curves defined over Q.

Theorem 3.1. $T(E, \mathbf{Q})$ is either cyclic of order ≤ 10, or order 12, or else is a direct product of a cyclic group of order 2 with a cyclic group of order $2, 4$ or 6. Moreover, for each of these structures there is a single rationally-parametrized one parameter family of elliptic curves with that type of torsion subgroup.

This, of course, is only over the field of rational numbers, $\mathbf{Q}$, but the natural profile of the question requires understanding torsion phenomena for *all* elliptic curves over any fixed number number field. Here we have some exciting results due to a number of people, Merel, Oesterlé, Parent, Kamienny, and very recent progress due to Maarten Derickx, Sheldon Kamienny, William Stein, Michael Stoll, and van der Hoej. And yet there remains quite a project (computational exploration, and–of course—theoretical as well) to be done.

Fix a positive integer d and let $P(d)$ be the *largest* prime number p such that there exists an elliptic curve (without CM; i.e., without 'extra' endomorphisms) defined over some number field of degree $\leq d$ over $\mathbf{Q}$ and for which there is a point of order p on that elliptic curve, rational over that field.

So, my theorem says that $P(1) = 7$. Only for small p is $P(d)$ known. Kamienny proved that $P(2) = 13$. It is also true (a forthcoming article of Derickx, Kamienny, and me) that the only examples of 13-torsion on elliptic curves over quadratic fields comes from a single rationally parametrized (infinite) family of them.

Parent, building on work of Kamienny, showed $P(3) = 13$, and recently Maarten Derickx, Sheldon Kamienny, William Stein, and Michael Stoll showed that $P(4) = 17$. Here it is similarly true that the only examples of 17-torsion on elliptic curves over quartic fields comes from three distinct rationally parametrized (infinite) families of them. This is contained in forthcoming joint work with Maarten Derickx and Sheldon Kamienny that focuses on the diophantine analysis of what we call *basic Brill-Noether modular varieties.*

One knows that $P(5) = 19$ and in all the cases I've just listed, all primes $\leq P(d)$ do occur as rational p-torsion for some elliptic curve defined over some field of degree $\leq d$.

But what about results for general values of d?

Here we have the deep theorem of Merel that for any d, $P(d) < \infty$. For a more specific upper bound, Merel's work with improvements from Oesterlé and Parent shows—for general d—that

$$P(d) \leq (1 + 3^{d/2})^2.$$

Or, to round it out,

$$P(d) << 3^d.$$

An exponential bound, in other words.

To gauge how close this upper bound comes to the actual phenomena, let's contemplate lower bounds. The trivial lower bound is

$$(*) \quad d^{1/2} << P(d),$$

and here's a proof of this. Take any elliptic curve E over $\mathbf{Q}$ and note that over $\bar{\mathbf{Q}}$, the algebraic closure of $\mathbf{Q}$, the kernel of multiplication by p in E is a (p,p)-type group, i.e., a two dimensional vector space over the prime field $\mathbf{F}_p$ and the Galois group $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$ acts on this vector space through a subgroup of the general linear group $\Gamma \subset \mathrm{GL}_2(\mathbf{F}_p)$. If you pass to an extension field $K/\mathbf{Q}$ such that the Galois group $\mathrm{Gal}(\bar{\mathbf{Q}}/K)$ acts through a subgroup Δ of triangular matrices of the form

$$\begin{bmatrix} 1 & * \\ * & * \end{bmatrix}$$

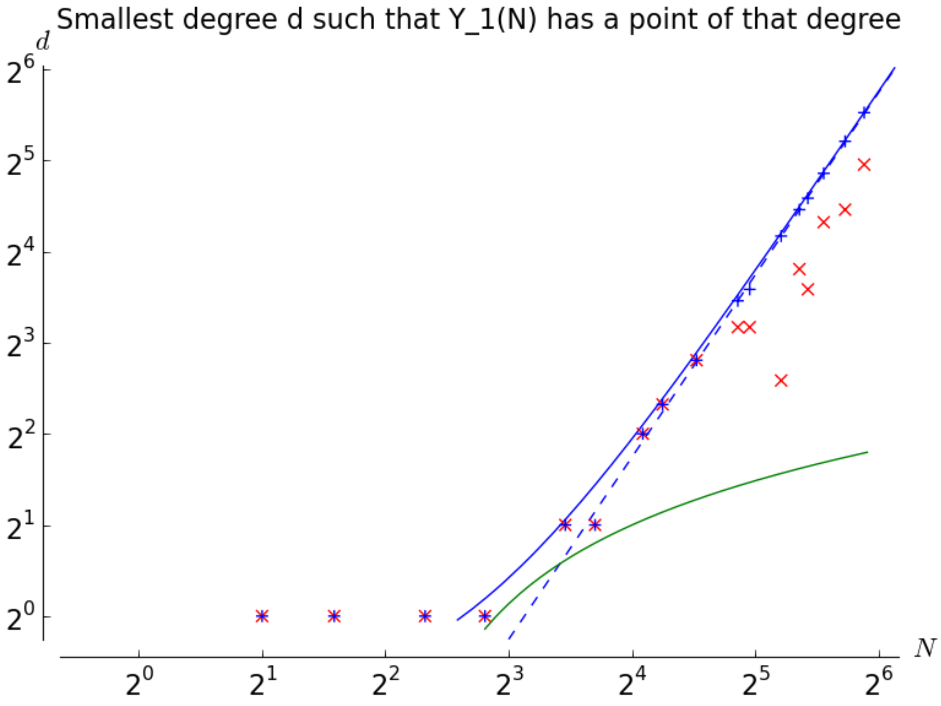
then E will have a torsion point of order p rational over this K. Since $[\mathrm{GL}_2(\mathbf{F}_p) : \Delta] = p^2 - 1) = O(p^2)$, the degree of such a K is $\leq p^2$, which gives **(*)**.

Note: A more geometric way of saying the same thing is to make use of the natural mapping—defined over $\mathbf{Q}$—of the modular curve $X_1(p)$ to the j-line. This is of degree $\frac{p^2-1}{2}$ observing that this gives a natural rationally parametrized family of elliptic curves with rational p-torsion over fields of that degree.

So we have

$$d^{1/2} << P(d) << 3^d.$$

Since no other wholesale construction of larger p-torsion in fields of degree d comes to mind, the minimalist instinct would then nudge one to consider the possibility that there would be a polynomial, rather than exponential upper bound for $P(d)$, andperhaps even an upper bound of the form $P(d) << d^{1/2+\epsilon}$. Here below is a graph computed by the first author of the present article jointly with Mark van Hoej. It is a log-log plot where the axes are $(x, y) = (\log p, \log d)$, the data points recording examples of 'lowest' degree d for the corresponding p occurs as prime torsion in a non-CM elliptic curve (over a field of degree d). The quotation-marks around the word 'lowest' is meant to signal that the blue data points and the blue extrapolated line corresponds to the lowest d for which there is a rational family of such examples of prime torsion p over fields of degree d. The red data points correspond to the sporadic points. The green curve is the proved (exponential) lower bound relating d to p. Visibly, much more computation needs to be done if we are to be able to surmise any general behavior with some feeling that there is evidence behind our guess.



In the literature, some conjectures give upper bounds for primes of torsion in elliptic curves of degree d, but since these published conjectures also consider prime torsion in CM elliptic curves, which our "$P(d)$" doesn't register, those conjectures necessarily must allow for an essentially linear lower bound[1].

There have indeed been conjectures approaching this.

Explicitly,

Conjecture 3.2. (Clark, Cook, J. Stankewicz)

$$P(d) \;<<\; d \log\log(d),$$

Conjecture 3.3. (Lozano-Robledo)

$$P(d) \;<<\; d.$$

It is tempting, then, to focus on the exponent of d related to the rate of increase of $P(d)$, i.e., to define:

[1] One might imagine distinctive bimodal behavior, for prime torsion in elliptic curves without CM over fields of degree d versus prime torsion in elliptic curves with CM.

$$e(d) := \frac{\log P(d)}{\log d}$$

and to ask:

Question 3.4. Can one find infinitely many values of d with $e(d)$ strictly greater than $\frac{1}{2}$?

4. Rank

4.1. **Density questions having to do with rank.** Let K be a fixed number field and consider the collection of all elliptic curves defined over K. The most natural 'first question' that is somewhat of a statistical nature that you might ask about Mordell-Weil rank is:

> Does $r(E; K)$ admit a finite upper bound (for fixed K and all elliptic curves over K)?

Here, far from actually having a resolution of this yes or no question, we don't even seem to enjoy a uniform consensus about guesses for what the truth is here, even for the field **Q**. (There are number theorists who believe yes, and others who believe no.) The following chart, which I got off the web, tabulates world's record large ranks for elliptic curves over **Q**—so far— with the year of their discovery and the winners.

rank $\geq$	year	Author(s)
3	1938	$Billing$
4	1945	$Wiman$
6	1974	$Penney - Pomerance$
7	1975	$Penney - Pomerance$
8	1977	$Grunewald - Zimmert$
9	1977	$Brumer - Kramer$
12	1982	$Mestre$
14	1986	$Mestre$
15	1992	$Mestre$
17	1992	$Nagao$
19	1992	$Fermigier$
20	1993	$Nagao$
21	1994	$Nagao - Kouya$
22	1997	$Fermigier$
23	1998	$Martin - McMillen$
24	2000	$Martin - McMillen$
28	2006	$Elkies$

Our knowledge, and the precision of our expectations, about densities, however, is somewhat more advanced.

4.2. **The computable upper bound, and the constraint of *parity*.**

- **A theorem:** For every prime number p there is a *computable* number $r_p(E,K)$—called the **reduced mod p-Selmer rank**—that constitutes an upper bound for the Mordell-Weil rank:

$$r(E,K) \leq r_p(E,K).$$

- **A Conjecture:**

$$r(E,K) \equiv r_p(E,K) \mod 2,$$

  i.e., the Mordell-Weil rank is of the same *parity* as the reduced mod p-Selmer rank (for every p).

- **A Fact:** We have (at least) the beginning of an understanding of statistical questions regarding the parity of reduced mod p-Selmer rank (and this conjecturally translates to a similar understanding of the statistics of Mordell-Weil rank).

Let's make some guesses now about rank, following the minimalist instinct. However, at this point it pays

- to repeat that *parity* is indeed a constraint and something that one must take careful account of, before making guesses, and
- to note that to do statistics about infinitely many instances one must say how one orders them. The ordering arrangement doesn't have to be a full linear ordering, but at the very least it should be given by an increasing system of finite subsets of the objects that are being studied, where the union of all these finite subsets is the whole. Then, one can talk about densities, or probabilities of features.

We will discuss statistics for the following two types of families.

(1) *All* elliptic curves defined over a fixed number field K. This infinite collection is "ordered" by the size of the absolute value of the norm of the conductor.

(2) *All* quadratic twists of a given elliptic curve E over a given field K. This boil down to considering the class of elliptic curves expressible by the equations

$$E^{(d)}: \quad dy^2 = x^3 + ax + b$$

for $a, b, d \in K$, with a, b fixed and d an integer of K, varying (mod squares). This infinite collection is "ordered" by the maximum size of the absolute value of the norm of any prime ideal dividing d.

The minimalist instinct then suggests:

Question 4.1. Is it true that, in either of these cases, if we consider the statistics of the sub-collection with *even* Mordell-Weil rank parity, it is 100% likely that the Mordell-Weil rank of a member of that family is 0? And as for the statistics of the sub-collection with *odd* Mordell-Weil rank parity, is it 100% likely that the Mordell-Weil rank of a member of that family is 1?

(For the second type of family, at least for those over $K = \mathbf{Q}$, this was already conjectured by Dorian Goldfeld in 1979.

Of course, to connect these expectations with a general sense of the average rank, we should either know or guess something about the density of parity.

4.3. **All elliptic curves over a fixed number field.** For the first type of family described above, i.e. for all elliptic curves defined over a fixed number field K, we expect that the distribution of even/odd parities is 50/50; i.e., half are even and half are odd, when the count is made according to the ordering that we described.

This would suggest the following target:

Conjecture 4.2. The average Mordell-Weil rank for *all* elliptic curves over any fixed number field K is $1/2$.

In 1992 Armand Brumer showed (by analytic means, and conditional on standard conjectures) that the average rank of elliptic curves over $K = \mathbf{Q}$ is bounded above by 2.3. More recently we have the magnificent achievement of Arul Shankar and Manjul Bhargava who established that it is bounded above by 0.99. This is by a formidable new tack on the geometry-of-numbers approach to counting mathematical objects related to this problem. Things are moving and we might hope for continued progress here in the coming years.

4.4. **Quadratic twist families.** Here we have some classical work by Heath-Brown for a specific family, and by Swinnerton-Dyer (with a recent improvement by Dan Kane) for the special case of elliptic curves over $\mathbf{Q}$ that have particular features related to their 4-torsion. Importantly, they establish finite average values of Mordell-Weil ranks for these families.

But, conceiving the problem for more general number fields one encounters a (surprising) new feature in the nature of parity itself. This is described in recent work of Zev Klagsbrun, Karl Rubin and myself. We deal with the mod 2-Selmer rank parity for a quadratic twist family over a number field K. This, then, is conjecturally the Mordell-Weil rank parity. We show that in the case where the number field K has at least one real embedding, the distribution of even/odd parities is 50/50. But even if you fix a specific elliptic curve E but allow your self to consider different choices of field K over which you gather parity statistics, the proportions of even/odd can change dramatically. For example, take the elliptic curve (labelled 50B1 by Cremona)

$$E: \quad y^2 + xy + y \ = \ x^3 + x^2 - 3x - 1.$$

By judicious choices of fields K one can obtain quadratic twist families whose mod 2-Selmer rank parity ratios take on a dense set of numbers in the range $(0, 1)$.

5. The future

is impossible to predict, but this is quite an exciting time and I expect that the Erdös spirit of conjecture and proof will continue to offer us the gift of marvelous mathematics.