Questions about Number

B. Mazur

(for the volume: New Directions in Mathematics)

If you read the chapter entitled **Proto-history** in André Weil's Number Theory : An approach through history From Hammurapi to Legendre , you might well be struck by how many of the earliest and the most innocent-sounding questions about numbers still hold much of their mystery for us today.

Not that there has been NO progress since the Babyloneans etched their cuneiform table of fifteen Pythagorean triples[1] or since Brahmagupta contemplated Pell's equation! It is rather the opposite course of events that has deepened the mystery for us: There has been progress. Questions about whole numbers have been studied with a range of powerful mathematical techniques; they have been illuminated by diverse mathematical structures.

And yet: we still seem to be novices, facing these questions.

The form of question-asking has evolved through the centuries. One might expect that the newer questions would get less "innocent", become more encrusted with theory, and stray further from the stuff of numbers that inspired Diophantus. But Mathematics has its inevitable, yet always surprising, way of returning to the simple. One simple, surely fundamental, question has been recently asked (by Masser and Oesterlé) as the distillation of some recent history of the subject, and of a good many ancient problems. This question is still unanswered, and goes under the name of the ABC-**Conjecture**. It has to do with the seemingly trite equation A + B + C = 0, but deals with this equation in a specially artful way.

---

[1] This tablet is labelled PLIMPTON 322 and dated to between 1900 and 1600 B.C., published in [N-S] ; See p. 9 of [We] for a photograph of it.

What might lead one to respect such an equation? We will examine this, and show how the "solutions" to this equation lace their way through a constellation of different mathematical structures all bearing on the nature of number, intermingling "Old Directions in Mathematics" with quite new ones. The discussion seems to break naturally into two parts, Part I requiring significantly less mathematical background than Part II. Two brief, but technical, synopses of proofs which come up in our discussion are given in Appendices A and B below. There also are a few "technical boxes" sprinkled at various points in the text which can perfectly well be skipped, but which treat peripheral issues which require more background than the text supposes.

For related expository reading, see the publications [Co], [Dar 2], [D-D-T], [Ed], [G], [H-R], [Ma 2], [Ri 1], and [R-S] cited in the bibliography below, and for further expository articles, consult the bibliography in [Ri 1].

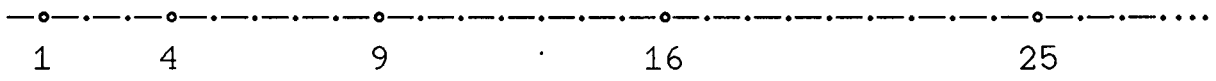
I am grateful to J. Cremona, H. Darmon, P. Diaconis, N. Elkies, F. Gouvea, A. Granville, R. Kaplan, K. Ribet, C. Stewart, and S. Wong for help, conversation, and comments about early drafts of this paper.

Part I

§1. Perfect powers.

Fibonnacci's treatise, Liber Quadratorum, written in 1225, is devoted to Diophantine questions about perfect squares. The prologue to it begins:

> "I thought about the origin of all square numbers and discovered that they arise out of the increasing sequence of odd numbers; for the unity is a square and from it is made the first square, namely 1; to this unity is added 3, making the second square, namely 4, with root 2; if the sum is added to the third odd number, namely 5,...." [Fi]



This manner of generating squares was already known to the Pythagoreans, and the similar recipe for generating cubes is described by Nicomachus in his Introduction to Arithmetic II:

> "For when the successive odd numbers are set out in an endless series beginning with unity, observe that the first makes the first cube, the sum of the next two makes the second cube, the sum of the next three following these makes the third cube, the sum of the four following these makes the fourth cube.... and so on indefinitely."

By a **perfect power** let us mean any power a^n of a whole number a, where the exponent n is greater than one. The arrays of perfect squares, perfect cubes, perfect fourth powers, etc. , (i.e., perfect n-th powers for each n=2,3,4,...) have long been a fount of innocent-sounding questions.

Looking at an array of perfect n-th powers on the number line, such as the array of squares pictured above, one is impressed only by the almost boring regularity in the spacing, and also by the thoroughly predictable way in which any two of these arrays "interact" (for example, the numbers that are both squares and cubes are precisely the perfect sixth powers).

But for any fixed n, m > 1 if *all we do* is to translate the array of perfect n-th powers (along the number line) by a fixed positive number k, i.e., by adding the integer k to each n-th power, and then if we ask:

Problem: *Determine the set of perfect* m*-th powers in the translated array* ;

or in other words,

Problem: *Find all solutions to the equation*

(1) $$X^m = Y^n + k$$

(*for* X *and* Y *natural numbers, and for the fixed exponents* n,m > 1 *and fixed positive integer* k),

we are in deep water.

Even simple specific instances of this problem can have quite surprising answers: for example, the reader might have difficulty guessing the *four* perfect cubes such that when 24 is added to each of them the results are perfect squares. That is, solve $X^2 = Y^3 + 24$ in integers X,Y, there being precisely four pairs (±X,Y) of solutions to this equation.

Answer: The four cubes that do the trick are 1, -8, 1000, and 542939080312. If we graph the equation $X^2 = Y^3 + 24$ in the Cartesian plane, as in Diagram 1 below, we can comfortably visualize the first three pairs of solutions (±X,Y) but to encompass the last pair on the same scale, this book would have to have a wing span of about twenty miles!

******* (put diagram 1 here)**********

The equation $X^2 = Y^3 + 24$

Diagram 1

The texture to this set of solutions to our problem posed above is not entirely untypical; that is: a cluster of solutions, plus one more solution which is noticeably larger than the rest, a sort of "top quark", and a good warning not to make conjectures about these matters on the basis of too limited numerical investigations. We will revisit this particular equation in §5 below.

The *qualitative* answer to the general **Problem** displayed above is that there are only a finite number of solutions to (1), i.e., only a finite number of m-th powers that are also in the array of n-th powers translated by the fixed positive number k.

This was known in 1929, by work of Siegel (following a line of development begun by Thue). The qualitative statement "only a finite number" is not much help, though, if, for some reason, you actually want to *find* the set of m-th powers in such a translated array. Nor does it help even if you have the somewhat less ambitious aim of giving an a priori upper bound for the size of the perfect m-th powers which are of the form: a perfect n-th power plus k. "A priori upper bound" here just means to give such an upper bound which is relatively easy to calculate, given the exponents n,m and the displacement k.

Some forty years after Siegel's Theorem was proved, the work of Baker (on lower bounds for nonvanishing linear forms in logarithms; cf. [Ba 3]) provided such a priori upper bounds. But even these upper bounds, sharpened in a series of papers [Ba 1,2], are not yet always sharp enough to closely reflect the thorny, and fascinating, numerical phenomena here.

To underscore the gap between qualitative results, and a thorough-going "finding of all solutions" for the type of problems such as the one displayed, consider this astounding result:

Theorem (Tijdeman, 1976): *There are at most a finite number of pairs of consecutive perfect powers .*

Or, in terms of equation (1), this theorem says:

Theorem : *Setting* k=1, *there are at most a finite number of solutions to the equation* (1)-- *even when one allows* m *and* n *to vary arbitrarily through all numbers greater than one.*

The status of the Diophantine problem posed by (1) in the case k =1 is quite special. In the case k=2, for example, or for *any* fixed k different from 1, we still do not even have a proof of finiteness of the number of solutions of (1) if m and n are allowed to range over all numbers > 1 (but such a finiteness statement would follow from the ABC-Conjecture below).

The proof of Tijdeman's Theorem depends upon the theory of lower bounds for nonvanishing linear forms in logarithms [T]; see [S-T] for a complete exposition of the proof; we will also give the briefest sketch of the main tactics of the proof in Appendix B below.

The only known example of a consecutive pair of perfect powers is 8 and 9. The general guess (made originally by Catalan in 1844; cf. Ribenboim's book, Catalan's Conjecture [Riben 2] for background) is that 8 and 9 is the *only* such example.

And here, even though Tijdeman's result actually assures us that there is a computable a priori upper bound to the size of such consecutive pairs of perfect powers, this computable bound is so high that we remain ignorant of whether or not this guess is correct. For an up-to-date description of where the current work is on this, see Baker's recent review [Ba 4] of [Riben 2]. In particular, I

understand from Baker's account that one knows the following facts about possible exponents m and n of consecutive perfect powers: both m and n must be at least 100, (due to work of Mignotte, using results of Inkeri) and the larger of the two exponents is known to be $\leq 10^{19}$, the smaller $\leq 10^{13}$ (results of Glass and coworkers at Bowling Green State University; see [G-M-O-S] and [L-M-N]). For more on the nature of explicit bounds in this problem cf. p. 217 of [S-T] (e.g., work of Langevin, elaborating Tijdeman's proof, provides an upper bound of

$$e^{e^{e^{730}}}$$

for any perfect power occurring in a pair of consecutive perfect powers).

The undisputed favorite among questions about perfect powers of numbers is **Fermat's Last Theorem** which asserts that the sum of two cubes is never a cube, the sum of two fourth powers is never a fourth power, and so on. The cleanness and simplicity of its statement, the pointed contrast of the behavior of higher powers with that of squares (for one has an infinity of instances of a sum of two relatively prime squares equal to a square), the enigmatic way the statement made its entrance onto the stage of Mathematics with a coy hint of the existence of a "marvelous proof", the way in which any proof of it, marvelous or not, failed to surface for three centuries, and the great amount of Mathematics its pursuit has given rise to, culminating in its recent splendid resolution (by Andrew Wiles, completed by Taylor-Wiles, [Wi], [T-W], using prior work of Frey, and the key "level-lowering theory" of Ribet which in turn was inspired by an important conjecture of Serre [S 2])-- all this justifies the special place this 17-th century question has held in the imagination of many people who think about

numbers.

§2. The "odds" of hitting on a solution.

But perhaps it is time to backtrack, to develop a bit of intuition, which might allow us to hazard guesses on which equations (such as among those displayed above) could be expected to have few solutions, and which could be expected to have many. The idea here is that, if n and m are large, there as so few n-th powers and so few m-th powers, that an "accident", such as a solution to (1), is simply very rare. Without trying to justify this kind of reasoning let us simply indulge in it, and see where it leads. And, for variety, let us modify the context a bit, by contemplating integer (i.e., whole number) solutions (in the variables X,Y,Z) to equations of the general form

$$(2) \qquad aX^{\alpha} + bY^{\beta} + cZ^{\gamma} = 0$$

where a,b,c are fixed (nonzero) integers, where for simplicity let us assume that no two of the three coefficients a,b,c have a common factor, and where the exponents α,β,γ are fixed positive numbers. This includes, for example, equations like those in Fermat's Last Theorem.

Imagine that we are going to look for solutions to (2) in the following mindless way. Fix a large positive number T, and simply "try out" *all* possible integer choices of X,Y,Z subject to the cut-offs

$$(3) \qquad |X| \le T^{1/\alpha}, \qquad |Y| \le T^{1/\beta}, \qquad |Z| \le T^{1/\gamma},$$

and let us also make the extra requirement that X,Y, and Z are "relatively prime" (meaning that these three numbers have no common factor larger than 1)[2]. With

this trial-and-error-strategy, we are guaranteed that the left-hand-side of (2), i.e., LHS= $aX^{\alpha} + bY^{\beta} + cZ^{\gamma}$ is less, in absolute value, than a fixed constant times T. "Fixed" here means simply that the constant depends only on the parameters a,b,c and not on our choice of T.

At this point let us imagine betting on each trial. To bet effectively, of course, you need some rough-and-ready way of computing the odds. And with our near total lack of knowledge, a natural first guess is that each of these trials is "random" in the sense that the values of our left-hand-side, LHS, are evenly distributed over the entire "possible range". In a word, we are going to guess that LHS hits any given value in its possible range equally-often. Since there is a constant times T possible values, the expectation of hitting 0, (with the thoroughly unjustifiable assumption we have just made) i.e., the "odds" of getting a solution to (2) from any given trial would then be a constant times 1/T. Since the number of trials allowed to us by our trimming-strategy (3) is roughly a constant times

$$T^{1/\alpha} \cdot T^{1/\beta} \cdot T^{1/\gamma} = T^{1/\alpha+1/\beta+1/\gamma},$$

our "expected payoff", i.e., the number of solutions to (2) we might benightedly hope to get from this procedure, is the number of trials times the expectation for any one trial, i.e., a constant times

---

[2] We take this precaution for otherwise, visibly "nonrandom" phenomena will swamp the data; for example, given any solution (x,y,z) you get infinitely many other ones by taking, e.g., $(x\cdot\lambda^{\beta\gamma}, y\cdot\lambda^{\alpha\gamma}, z\cdot\lambda^{\alpha\beta})$ for integer values of λ. But also there is some subtler "nonrandom" behavior ruled out by our precaution; here is an example pointed out to me by Granville: consider the two-parameter family of rational solutions to the equation $X^3+Y^3=Z^4$ given by: $X= \lambda\cdot(\lambda^3+\mu^3)$, $Y= \mu\cdot(\lambda^3+\mu^3)$, $Z= \lambda^3+\mu^3$ parametrized by λ and μ.

(4) $T^{1/\alpha+1/\beta+1/\gamma} \cdot (1/T) = T^{(1/\alpha+1/\beta+1/\gamma\ -1)}$.

Let us now glance at the exponent in (4) in order to predict something about the qualitative behavior of the solutions to (3).

I. If $1/\alpha+1/\beta+1/\gamma$ is less than 1, the exponent in (4) is negative, so one might expect few solutions! And this is the case, as Darmon and Granville have recently proved[3] (cf. [D-G]):

Theorem (Darmon-Granville): *Let* a,b,c *be nonzero constants, no two of which have a common factor, and let* (α,β,γ) *satisfy the inequality*

$$1/\alpha+1/\beta+1/\gamma < 1.$$

Then there are only a finite number of solutions to the equation

$$aX^{\alpha} + bY^{\beta} + cZ^{\gamma} = 0$$

in (nonzero) triples of integers (X,Y,Z) *such that* X, Y, *and* Z *have no common factors.*

II. If $1/\alpha+1/\beta+1/\gamma$ is greater than 1, i.e., if (α,β,γ) written in nondecreasing order is among the entries of the table:

[3] by applying Faltings' theorem judiciously to Galois coverings of the projective line with ramification-signature (α,β,γ); see [D-G] for this elegant argument.

α	β	γ
1	*	*
2	2	*
2	3	3
2	3	4
2	3	5

(where * means any integer allowed by the convention of nondecreasing order),

the exponent in (4) is positive, so we might not be surprised to find that the equation has an infinity of solutions. This, of course, is subject to some sort of caveat, for there are, at times, visible facts about a particular equation (like a,b,c positive and α,β,γ even) that would preclude the equation from having too many solutions.

For each triple (α,β,γ) in the table above, it isn't hard to produce equations whose exponents are given by that triple, and which has an infinity of solutions in relatively prime integers (X,Y,Z); for example: the equation

$$X^{\alpha} + Y^{\beta} - Z^{\gamma} = 0 \qquad (5)$$

has this property, for any triple (α,β,γ) occurring in our table. For explicit "rationally parametrized" formulas for the infinitude of solutions in each of these cases, see [D-G]. The case $(\alpha,\beta,\gamma) = (2,3,5)$ is particularly interesting and I understand that F. Beukers has recently found the complete set of rationally parametrized families of infinite solutions to it (there are twenty-three such families). The case of $\alpha=\beta=2$ and arbitrary γ is simple, and ancient: for variables U,V, let X(U,V) and Y(U,V) be the homogeneous polynomials in U and V which come about as the real and imaginary terms of the expansion of the γ-th power of $(U+\sqrt{-1}\cdot V)$:

$$X(U,V) + \sqrt{-1}\cdot Y(U,V) = (U+\sqrt{-1}\cdot V)^{\gamma}.$$

Now multiply left and right side of the displayed equation by their respective complex conjugate to get

$$X(U,V)^2 + Y(U,V)^2 = (U^2+V^2)^{\gamma},$$

so that any substitution for (U,V) of a pair of relatively prime integers (u,v) gives the solution

$X=X(u,v)$, $Y= Y(u,v)$, and $Z = u^2+v^2$

to equation (5) with triple of exponents $(\alpha,\beta,\gamma) = (2,2,\gamma)$.

III. If $1/\alpha+1/\beta+1/\gamma$ is exactly 1, i.e., if (α,β,γ) is one of the three triples

α	β	γ
2	3	6
2	4	4
3	3	3,

the exponent in (4) is zero, so perhaps we had better hedge our bets. Hedging bets seems to be a good idea, in view of some numerical calculations that have been carried out with equations of exponents (α,β,γ) occurring in this table; e.g., consider the equation

E(m): $$X^3 + Y^3 + m\cdot Z^3 = 0.$$

Kramarz and Zagier [Z-K] have shown (making use of standard conjectures and computer calculations) that for precisely 10,292 of the cube-free numbers m in the range $1 < m < 20{,}000$, the equation E(m) has an infinitude of solutions (X,Y,Z) in integers with no common factors; i.e., for this range of coefficients m, roughly 62% of these equations have an infinity of relatively prime solutions, while the remaining 38% of them have only a finite

number of such solutions. See also the report in [G-P-Z] of more recent calculations carried out for all values of m in the range $|m| < 100{,}000$. If one restricts attention to equations E(m) where m is the negative of a prime number ($m = -p$) then we have somewhat more precise information: If $p \equiv 2,3$ or 5 modulo 9 there are no nontrivial solutions to E(m). If $p \equiv 4$ or 7 modulo 9, Elkies has recently announced that he can prove that there is an infinitude of solutions (x,y,z) with x,y, and z relatively prime. This leaves $p \equiv \pm 1$ modulo 9. When $p \equiv -1$ modulo 9 we *expect* that there is an infinitude of solutions

(again with x,y, and z relatively prime)[4], and, finally, the case $p \equiv 1$ modulo 9 is the interestingly erratic case: things can go either way; there are either no nontrivial solutions, or there is an infinitude of solutions (with x,y, and z relatively prime). For an account of all this, see [V-Z].

For further discussion of equations in all three categories I,II, III, see [D-G].

The trichotomy that we have fallen onto by this gambler's type of reasoning, i.e.,

$$1/\alpha+1/\beta+1/\gamma \; < \; 1$$

$$(6) \qquad 1/\alpha+1/\beta+1/\gamma \; = \; 1$$

$$1/\alpha+1/\beta+1/\gamma \; > \; 1,$$

is hardly a spurious one. It separates equations such as (2) above into three classes and this same three-way distinction can be rediscovered by considering the differential-geometric features of the locus of complex zeroes of these equations, or their algebraic geometric

[4] This would follow from the Conjecture of Birch and Swinnerton-Dyer

features, or even, to some extent, their topology.

I have often wondered what historical role this type of unjustified "probabilistic reasoning" has played in the shaping of mathematical subjects. Are these heuristic arguments used more as a predictive tool (as a guide for the establishment of some theory) or more as a mnemonic, or handy codification after some theory has been established? Whenever such a heuristic argument actually "works", i.e., conforms to theory or computation, we may derive from it, at least, some sense (or hope) that the analysis that went into it does not leave out any of the grosser features of the phenomena being studied. Number Theory has its share of these heuristic devices, some as perfectly explicit as our "gambler's argument" above, and others which are vaguer, but which still illuminate. It might repay the effort for a historian of Mathematics to examine these in a detailed scholarly way. There is the famous elementary calculation (by Gauss? and others?) giving the estimate of $1/\log x$ for the probability that a number "around the size of x" be prime, and leading to the conjecture (a version of the "Prime Number Theorem") that the number of primes $\leq x$ is asymptotic to

$$\int_2^x d\xi/\log \xi,$$

this being a visibly predictive use of such heuristics, at least in the sense that this result was eventually established much later, not in the lifetime of the original conjecturer(s). Nowadays, there are innumerable predictions and codifications in the subject which depend on some probabilistic model. For example, there is Montgomery's Conjecture that the gaps between the zeroes of the Riemann Zeta-function are distributed like the gaps between eigenvalues of large random Hermitian matrices[5]; there is the so-called "Cohen-Lenstra

Heuristics" which predict statistics on the structure of the ideal class group of quadratic fields; there is a significant elaboration of the very "gambler's reasoning" we described above which "predicts" the asymptotics of the number of rational points of "height" $\leq x$ on certain projective algebraic varieties (Manin's Conjecture).

Getting back to our particular subject, the reader might wonder, as I do, whether one can refine our gambler's heuristics so as to comment intelligently, one way or the other, on the statistical likelihood of the finiteness result given by Tijdeman's Theorem formulated above.

§3. ABC.

The reader might also ask for a more fluid context than is given by equations of the particular form of (2) in §2, for a discussion about powers of whole numbers and their relations. It seems amazing that despite many centuries of devotion to such questions, it was less than ten years ago that Mathematicians (specifically, Masser [Mas] and Oesterlé [Oe], refining an idea of Szpiro, and guided by a result about polynomial algebras due to Mason) formulated a startlingly simple problem that focuses on such a fluid context, and that still captures something of the essence of the type of question posed by equations of the form (2).

Masser and Oesterlé consider the humble *linear* equation

$$A + B + C = 0. \tag{7}$$

They boldly define an ABC-**solution** to be absolutely any

[5] For numerical work on this, see Odlyzko [Od], and for work on higher correlations of zeroes of the Riemann Zeta-function, see [Rud-Sar]

solution to (7) in relatively prime nonzero integers A,B,C. There is no obstruction, then, to finding as many ABC-solutions as you might want! The sensible tactic, though, is to sort through ABC-solutions "grading" them according to "interest", where an ABC-solution is considered "interesting" if A, B, C are divisible by high perfect powers. We will do this "grading" in a moment, but the guiding idea will then be to conjecture that there are relatively few "interesting" ABC-solutions; i.e., once you put a linear relation (7) on three relatively prime integers, Masser and Oesterlé will be conjecturing that there is a strong compulsion for these integers not to be highly divisible by perfect powers, where the adverb "highly" is about to be given a quantitative meaning.

If N is a nonzero number, define its **radical** to be that number which is the product of each of the distinct primes dividing N. Denote the radical of N by rad(N). So, for example: rad(12) = rad (18) = 6, and rad(2^{100}) = 2. Our point of view will be to think of a number N as being "highly divisible by perfect powers" if it is, roughly speaking, large in comparison with its radical.

Let us convene, for our ABC-solutions, to have C be the maximum of the three numbers A,B,C, in absolute value. By the **power** P of an ABC-solution (A,B,C) let us then mean the quantity:

$$P(A,B,C) = \log |C| \,/\, \log(\mathrm{rad}(A \cdot B \cdot C)).$$

If the power P of an ABC-solution is high, we want to think of that solution as being "highly divisible by powers". To check quickly that this is not an unreasonable way of thinking of P let us do the exercise of estimating P for an ABC-solution consisting of perfect n-th powers

$$A = a^n , \quad B = b^n , \quad C = c^n , \text{ for some } n;$$

i.e., the triple (a,b,c) would then be a nontrivial solution

to the Fermat equation of exponent n (which, of course, we now know does not exist for n >2, but let us follow through the consequences of the existence of such a solution).

Since $\log |C| = \log \max(|A|, |B|, |C|) \geq 1/3 \cdot \log(|A \cdot B \cdot C|)$

$$\geq n/3 \cdot \log(|a \cdot b \cdot c|) \quad \geq \quad n/3 \cdot \log(\mathrm{rad}(A \cdot B \cdot C)),$$

the "P" of such an ABC-solution would be $\geq n/3$, and hence would be large if n is large.

We are now ready for the formulation of the rather remarkable (and still unsolved!)

ABC Conjecture (Masser-Osterlé): *For any number* $\eta > 1$, *only a finite number of* ABC-*solutions can have power* $P \geq \eta$.

The beauty of such a Conjecture is that it captures the intuitive sense that triples of numbers which satisfy a linear relation, and which are divisible by high perfect powers, are rare; the precision of the Conjecture goads one to investigate this rarity quantitatively. Its very statement makes an attractive appeal to perform a range of numerical experiments that would test the empirical waters. On a theoretical level, it is enlightening to understand its relationship to the constellation of standard arithmetic theorems, conjectures, questions, etc., and we shall give some indications of this below. There is also the lure of actually trying to prove this conjecture, and if not the conjecture in its full strength, then perhaps something (even if a good deal weaker) in its direction. To give an example of such a weaker but more tractable statement (e.g. (9) below), first note that the ABC-Conjecture implies that there is a constant κ such that

(8) $$\log |C| < \kappa \cdot \log \mathrm{rad}(A \cdot B \cdot C)$$

for all ABC-solutions (A,B,C). This is because the ABC-Conjecture implies that for any fixed number η greater than 1, there is only a finite set $S(\eta)$ of ABC-solutions with power $P > \eta$; so fix such an η and take κ greater than the maximum power of all ABC-solutions in $S(\eta)$.

Now since (8) may be out of reach at present, and since $\log \mathrm{rad}(A\cdot B\cdot C)$ goes to infinity more slowly than any fixed positive power of $\mathrm{rad}(A\cdot B\cdot C)$, one might try to establish an inequality of the form:

$$(9) \qquad \log |C| < \kappa \cdot \mathrm{rad}(A\cdot B\cdot C)^{\delta},$$

valid for all ABC-solutions, for δ some fixed positive number, as a gauge of how powerful the available methods are: the smaller δ one can prove this for, the better. In fact, Baker's "theory of lower bounds on linear forms in logarithms" implies such inequalities. At the present time[6] this inequality is known for any exponent $\delta > 2/3$, where the constant κ, dependent upon δ, is effectively computable.

In the direction of making further conjectures, the bluntly qualitative form of the conjecture ("*only a finite number of* ABC-*solutions*") begs to be sharpened to more precise quantitative statements (e.g, it would be good to have a conjecture carrying some conviction, that gives an explicit upper bound for $|A\cdot B\cdot C|$ for ABC-solutions such that $P > \eta > 1$, as a function of η).

As for numerical experiments, no one, to my knowledge, has yet found an ABC-solution with $P \geq 2$. And note that by the exercise we did above, a *proof* that there are no ABC-solutions with $P \geq 2$ would give another proof of

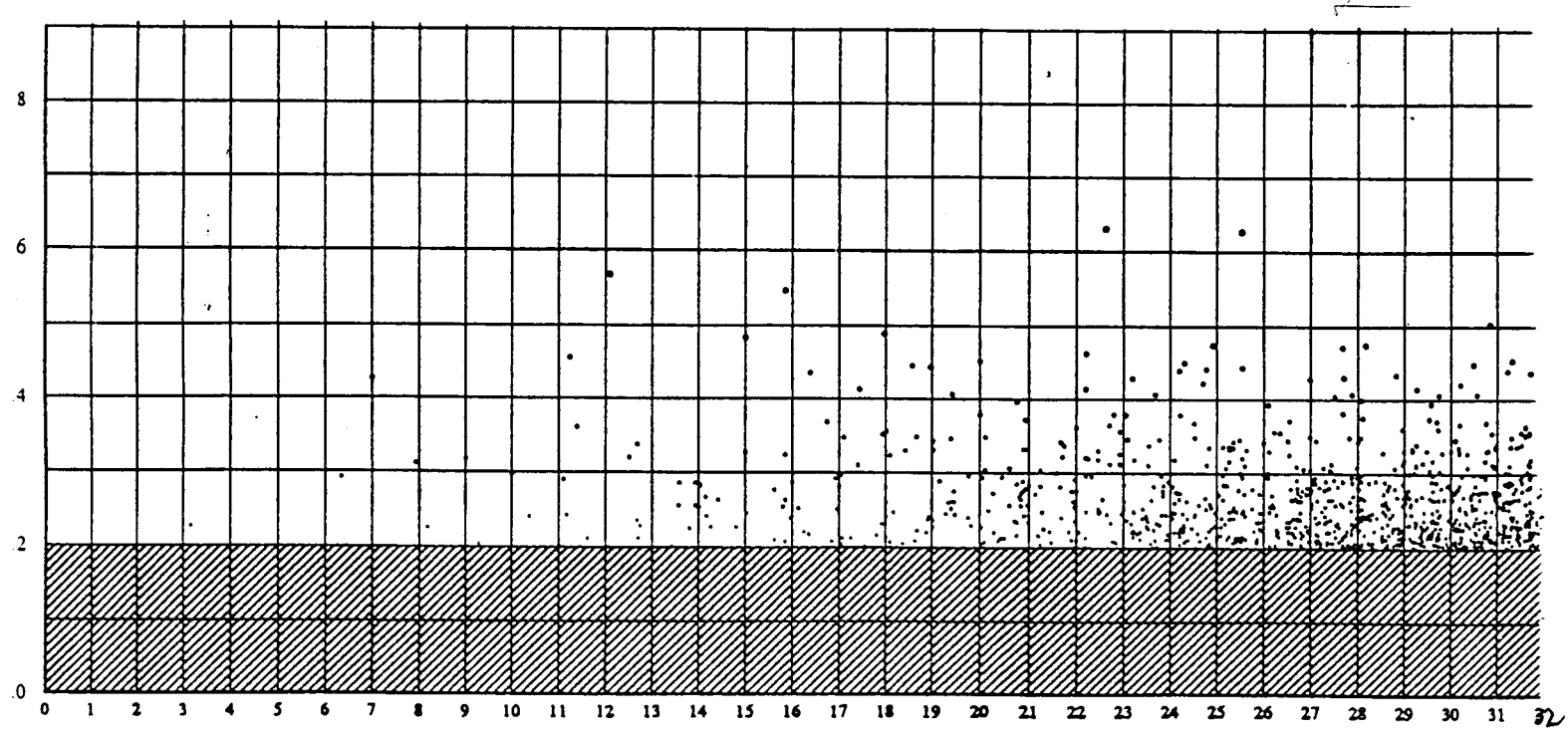
[6] See [S-Y]. This is an improvement of a prior inequality due to Stewart and Tijdeman and incorporates ideas of Waldschmidt

Fermat's Last Theorem for exponents > 5. As a similar exercise, using only elementary algebra, it is easy to show that the ABC-Conjecture (even without any particular upper bound given for P) would imply the Theorem of Darmon-Granville quoted above, as well.

The four most "powerful" ABC-solutions presently known, taken from a table in [B-B] , are:

	Equation	P
1.	$2 + 3^{10}.109 + (-23^5) = 0$	1.629912
2.	$11^2 + 3^2.5^6.7^3 + (-2^{21}.23) = 0$	1.625991
3.	$283 + 5^{11}.13^2 + (-2^8.3^8.17^3) = 0$	1.580756
4.	$1 + 2.3^7 + (-5^4.7) = 0$	1.567887

These examples were discovered by the mathematicians Reyssat, de Weger, Browkin-Brzeziński, and de Weger, respectively. Elkies and Kanapka have systematically tabulated all ABC-solutions (where $|A| \leq |B| \leq |C|$) whose power is greater than 1.2, in the range $|C| < 2^{32}$. There are 986 such ABC-solutions, and this tabulation is displayed by the printing of a dot with x,y coordinates ($\log_2 |C|$, P) for each such ABC-solution:



§4. Digression on ABC and Mordell's Conjecture.

There is a direct theoretical connection between the ABC-conjecture and some of the more classical problems in arithmetic, besides the connection that we have already seen between ABC and Fermat's Last Theorem. In this section, *which can be skipped in that it will not be referred to later in this article* , I want to give a brief description of Mordell's Conjecture because of its immense importance to our subject, and also because Elkies has shown by a fairly elementary argument that the ABC-Conjecture (for number fields) implies the Mordell Conjecture (see [E]; we will sketch this argument in appendix A below).

The Mordell Conjecture was originally formulated in 1922, and it was first proved by Faltings in 1983. It is about *rational* solutions (x,y) of polynomial equations P(X,Y)=0. That is, the problem it addresses is the study of pairs of

rational numbers x, and y such that P(x,y)=0. This is in contrast to the type of question we have been asking so far where the focus has been rather on *integer solutions* . Although these two kinds of problems, to find *rational* or to find *integral* solutions, are visibly related, there are many qualitative differences between them. I will return to one somewhat surprising difference at the end of this section.

The full assertion of Faltings' Theorem (Mordell's Conjecture) in technical language asserts that any algebraic curve of genus > 1 over any algebraic number field has at most a finite number of rational points. For an introductory discussion of the notions of **algebraic curve** and **genus**, and of Faltings' Theorem, see [Ma 2].

We can illustrate the power of Falting's Theorem by considering this example which can be stated in completely elementary terms[7]. Fix n an integer ≥ 5. Let G(X) be any polynomial of degree n,

$$G(X) = X^n + a_{n-1}\cdot X^{n-1} + a_{n-2}\cdot X^{n-2} + \ldots a_0$$

with coefficients a_j which are rational numbers, and such that G(X) has no "multiple roots" when it is factored over the complex numbers. A convenient necessary and sufficient criterion for G to have no multiple roots is that the polynomial G(X) and its derivative G'(X) have greatest common divisor equal to 1. It follows from Faltings' Theorem that the equation

$$Y^2 = G(X)$$

has at most a *finite number* of rational solutions (x,y). If you wish, another way of saying this is that as you allow

[7] "stated in elementary terms", yes, but definitely not proved by elementary means!

x to run through *all* rational numbers, the values G(x) are almost never *squares* of rational numbers, and more precisely they are squares for at most a finite number of choices of x.

While we are considering this example, we might ask *how many* rational solutions can an equation of the form $Y^2=G(X)$ have? Recently it has become (at least) plausible to hope that this number is bounded only by the degree of G. Specifically

Conjecture: For each $n \geq 5$ there is a number $B(n) < \infty$, such that for any polynomial G(X) of degree n with no multiple roots, the equation

$$Y^2 = G(X)$$

has no more than B(n) rational solutions.

For reasons for this to be plausible, see [C-H-M 1,2]. From an experimental point of view, it seems to be hard to come up with polynomials G of small degree ≥ 5 (say, precisely of degree 5) for which the displayed equation above has a large quantity of rational solutions. As I am writing this, the record (for polynomials of degree 5) is held by Kulesz and Keller [K-K]: they have found an example having ≥ 588 points. But we still lack sufficient experience here to even begin to guess whether this is close to optimal or very far from it (E.g., is the maximum number of solutions for polynomials G of degree 5 on the order of 10^3? Or is it closer to 10^{10^3}?) Or is the above Conjecture false and is there no uniform bound at all?

One "surprising difference" between rational vs. integral questions is in our present understanding of "decidability issues" related to these questions. Well over two decades ago, Matijasevic explicitly produced a

polynomial $P(T; X_1,...,X_m)$ in the variables T and the X's with integral coefficients for which there does NOT exist a computer program which, for any given specialization of the variable T, $T \mapsto 1$, $T \mapsto 2$, $T \mapsto 3$,.... and in general $T \mapsto t_o$, correctly answers the question of whether or not the polynomial equation

$$P(t_o; X_1,...,X_m) = 0$$

has an integral solution in the variables X_j. In a word, the problem of deciding whether or not a given polynomial has integral solutions is "unsolvable". But, to this day, one does not know whether the corresponding problem for *rational* solutions is decidable!

§5. The passage from ABC to cubic curves. Nothing could be simpler. Given an ABC-solution

$$(\varepsilon) \qquad A + B + C = 0,$$

(recall that A,B,C are integers with no common factors) you write the cubic equation

$$E_{(\varepsilon)}: \qquad Y^2 = X\cdot(X-A)\cdot(X+B).$$

The intended effect of writing such an equation is to invoke its locus of (say, complex-valued) zeroes, i.e., pairs of complex numbers (x,y) such that $y^2 = x\cdot(x-A)\cdot(x+B)$. These points (x,y) on $E_{(\varepsilon)}$ trace out a smooth plane cubic curve in (X,Y) space. If we were to complete (X,Y)-space to form the projective plane (by adding a line "at infinity") our curve $E_{(\varepsilon)}$ would have one extra point "*at infinity*" and in the discussion below we include that extra point (denoted 0) in the locus $E_{(\varepsilon)}$.

Let us review the geometric construction which provides an extremely important *addition law* on the points of $E_{(\varepsilon)}$. That is, given any two points u, v of $E_{(\varepsilon)}$, we will define a point which we will call u + v in $E_{(\varepsilon)}$. The reason for using the + sign here is to signal that this operation is an "addition law" in the sense that it satisfies the usual laws that addition in arithmetic satisfies. Explicitly: this addition law is commutative and associative; the point of $E_{(\varepsilon)}$ referred to as "0" above plays the role of "zero-element" in the sense that 0 + u = u for any point u of $E_{(\varepsilon)}$; and given any point u of $E_{(\varepsilon)}$ there is an "additive inverse" which we might call -u with the property that u + (-u) = 0. In other words, the set of points of $E_{(\varepsilon)}$ with this operation "+" forms a *commutative group*. The key fact that allows us to define such a law of addition is that any straight line ℓ in the (X,Y)-plane intersects a cubic curve $E_{(\varepsilon)}$ in precisely *three points*. That is, this will be true if we interpret things correctly! For a number of things may seem to conspire to make that statement false. First, if our straight line is tangent to $E_{(\varepsilon)}$ at some point u we have must interpret u as being a *double* point of intersection of ℓ and $E_{(\varepsilon)}$. Second, we must not only count intersection points (x,y) with x and y real numbers for then we might miss some intersection points: we must allow x and y to be complex as well. Third, we must not forget that the "extra point at infinity" on $E_{(q)}$ which we have labelled 0 may very well occur as an intersection point: specifically a line ℓ in the (X,Y)-plane contains 0 if and only if it is vertical.

With all these provisos, a characterizing property of this law of addition is that any three points u, v, w on $E_{(\varepsilon)}$ which lie on a line in the (X,Y)-plane sum up to 0.

It follows from this characterizing property that if w =(x,y) is a point of $E_{(\varepsilon)}$, then its inverse, -w, ("additive

inverse" in the sense of this addition law on $E_{(\varepsilon)}$) is the point (x,-y). To see this from the above discussion, draw the vertical line ℓ in the (X,Y)-plane passing through w. Since ℓ passes through the point 0 as well, the third point of intersection w' in $\ell \cap E_{(\varepsilon)}$, which is visibly the point (x,-y), has the property that w + 0 + w' sums to 0, i.e., w' is an additive inverse to w.

Since any two distinct points u,v on the curve $E_{(\varepsilon)}$ determine a unique line ℓ in (X,Y)-space going through them (ℓ = the "chord" passing through u and v) and this chord ℓ has a unique third point of intersection (call it w) with our cubic curve $E_{(\varepsilon)}$:-- our recipe gives u + v = -w.

**********(put diagram 2 here)**********

The addition law of points on $Y^2 = X^3 - X$.

Diagram 2

It is natural (and rather forced on us) to define u + u to be -w, where w is the third point of intersection of the curve $E_{(\varepsilon)}$ with the unique line ℓ tangent to $E_{(\varepsilon)}$ at the point u:

**********(put diagram 3 here)**********

Twice a point in $Y^2 = X^3 - X$

Diagram 3

From its very description, it is clear that this law (for the "addition" of points on $E_{(\varepsilon)}$) is commutative; the fact that this law is also *associative,* is proved by fairly elementary means but is, nevertheless, a minor miracle, which has been rediscovered in different ways, and put to

different uses, over the course of centuries. If you haven't seen this proved before, it would repay the effort to do this: Construct the "triple sums" (u+v)+ w and u+(v+ w) in Diagram 4 below by simply drawing the appropriate straight lines on that diagram to construct in turn the points u+v, v+ w, and then the triple sums to check, by eye, that these triple sums are in fact equal. This, of course, is **not a proof**. But this exercise already gives a sense of what sort of statement in Plane Projective Geometry it is, to affirm that these two triple sums are equal.

**********(put diagram 4 here)**********

The associative law on $Y^2 = X^3 - X$

Diagram 4

An algebraic curve such as $E_{(\varepsilon)}$ (e.g., any plane cubic curve $y^2 = g(x)$ where g(x) is a cubic polynomial in x with no multiple roots) together with this attendant additive law for its points, is called an **elliptic curve.**

This additive law for any given $E_{(\varepsilon)}$ has the convenient aspect of being "algebraic" in the sense that the coordinates of u + v may be given in terms of rational functions of the coordinates of u and of v; for example, as an exercise in the definition of the "addition law" plus a bit of plane geometry, you can try to derive the formula for the coordinates of u + u in terms of the coordinates of u.

For a short account of elliptic curves see [G].

The elliptic curve $E_{(\varepsilon)}$ is often referred to as the **Frey curve** of the ABC-solution (ε) in honor of Gerhart Frey, who realized that there is a distinct advantage to

changing the focus of our attention from the ABC-solution (ε) to this elliptic curve $E_{(\varepsilon)}$, especially if we are interested in ABC-solutions of high power P; See [Fr] and the prior, closely related, construction due to Hellegouarch in the early 70's given, e.g., in the discussion preceding Th. 4 in [He]). Frey noticed that one can re-express the hypothesis that an ABC-solution (ε) has the property that $A \cdot B \cdot C$ is divisible by a perfect power, as a specific, and sometimes quite "telling", property of the group structure of the elliptic curve $E_{(\varepsilon)}$. Roughly speaking, the more divisible by perfect powers the ABC-solution (ε) is, the more *peculiar* the corresponding Frey curve $E_{(\varepsilon)}$ is. What it means for an elliptic curve to be *peculiar* , however, we must leave for Part II below. The point is that we do have almost a century's worth of detailed mathematical theory concerning the arithmetic of elliptic curves, giving us a fairly developed sense of what to expect, and what not to expect, in the way of elliptic curves and their arithmetic behavior. And if we start with an ABC-solution (ε) where A, B, and C are perfect n-th powers for n a prime number ≥ 5 (i.e., a solution of Fermat's Last Theorem for prime exponent ≥ 5) the corresponding Frey curve $E_{(\varepsilon)}$ seemed so peculiar, that no one working in the field thought that such an elliptic curve could plausibly exist. Plausible or not, though, its actual existence could not be ruled out until the recent advance due to Wiles and Taylor-Wiles. We will be giving brief hints below; for more elaborate and excellent accounts of this story, see [Co], [Dar 2], [D-D-T], [G], [Ri 1], and [R-S].

But let us take a step backwards and ask what kind of a thing we are doing when we make a "transformation" such as:

(8) ABC-solution (ε) $A+B+C = 0$ -----> Elliptic curve $E_{(\varepsilon)} : Y^2 = X \cdot (X-A) \cdot (X+B)$.

Ignoring its specifics, this "transformation" is in the format of

(9)	The set of all solutions to a certain equation EQU.	------->	The set of all examples of a specific mathematical structure STR.

Now it is often a healthy sign, in studying an equation, if you find yourself dealing with such a format. There are two clear reasons to be pleased when this happens: First, if you have established a rule such as (9), then every time you have a solution to your equation EQU, you don't only have a solution to a particular equation, you have something more: you have "animated" the solution by relating it to a specific instance of the mathematical structure STR, which has, perhaps, interesting features of its own, and may be worth further study in its own right. But going the other way, by understanding conceptually, and perhaps classifying the structures STR, you might gain a new technique for constructing, or constricting, or just understanding better the solutions to your equation EQU.

Sometimes such a transformation as (9) helps in simply counting structure, or solutions to equations; we will consider this question of counting, with regard to the transformation (8), in the box labelled [*1].

And sometimes it is useful to study transformations that go in the other direction, from "structures" to solutions of an equation.[8]

[8] The example of this most closely like an "inverse" to the transformation

********(put box labelled [*1] here)**********

§6. The Mordell Equations.

Consider the integer solutions (X,Y) of the equation

(10) $X^2 = Y^3 + k$

for some fixed non-zero integer k. These equations are special cases of equation (1) discussed in §1, i.e., we have fixed the exponents m,n in equation (1) to be m=2 and n=3. The study of the system of equations (10) for k = ±1,±2, ... occupies a position in the history of Diophantine equations somewhat akin to the position that the study of fruitflies occupies in genetics: these are intensely studied "model systems". The equations (10), called **Mordell's Equations**, have an extensive literature, and constitute a showcase for the various methods that can be brought to bear on similar problems. A particular attraction of the Mordell equations is that they are connected to the theory of elliptic curves in at least two (somewhat incommensurate ways). For one thing, for each k, the Mordell equation (being of degree 3) is the equation of an elliptic curve. For another, the Mordell equations provide us with another illustrative example of the sort of transformation (9) that we talked about in the previous section: For any $k = 1728 \cdot \Delta$ *each rational solution* (b,a) of the Mordell equation

(8) is given by the classical theory of *moduli* for elliptic curves. This classical theory constructs a natural transformation that passes, e.g., from pairs consisting of an elliptic curve together with a chosen cyclic subgroup of order N in it, to solutions of a specific polynomial equation in two variables, the "modular equation" of level N, or, essentially equivalently, to points on a specific algebraic curve, the "modular curve" $X_0(N)$.

determines an elliptic curve $E_{(a,b)}$ given by a cubic equation of discriminant equal to Δ:

$$E_{(a,b)}: \quad y^2 = x^3 - (a/48)\cdot x - (b/864).$$

The integral points (X,Y) of Mordell's Equations (10) are entirely known for $|k| \leq 10{,}000$ and known with the exception of about 1000 values of k for $|k| \leq 100{,}000$ (these computations are very recent; cf. [G-P-Z]). A Conjecture due to M. Hall asserts that the integral solutions are bounded by the size of k according to the following rule:

Hall's Conjecture: There is a constant C such that

$$|Y|^{1/2} < C \cdot |k|.$$

Given any integer solution (X,Y) of equation (10) for some k, the ratio $|Y|^{1/2}/|k|$, then, gives us a *lower bound* for the constant C conjectured to exist by Hall. For example, the largest integral point on the curve

$$(11) \quad X^2 = Y^3 + 24$$

which we discussed in §1 has its Y-coordinate equal to 8158 and therefore the ratio $|Y|^{1/2}/|k|$ is 3.76...

The data of [G-P-Z] suggests that C might indeed be relatively small: the largest value for this ratio $|Y|^{1/2}/|k|$ achieved by *any* integral point (X,Y) of (10) that Gebel, Petho, and Zimmer find (in the range $|k| \leq 100{,}000$) is 4.87... and, in fact, all the integral points tabulated in [G-P-Z] with ratio $|Y|^{1/2}/|k|$ greater than 1.5 are given in the following table.

Table of some large integral points
(taken from [G-P-Z])

k	Y	$\lvert Y\rvert^{1/2}/\lvert k\rvert$
17	5,234	4.26...
24	8,158	3.76...
-207	367,806	2.93...
225	720,114	3.77...
-307	939,787	3.16...
1,090	28,187,351	4.87...
28,024	3,790,689,201	2.20...

Assuming Hall's Conjecture, one can define another constant, call it c, which is relevant to the above data. Namely,

$$c = \text{lim.sup. } |Y|^{1/2}/|k|$$

where the "lim. sup" is taken over all integral pairs (X,Y) with $k := X^2 - Y^3$. That is, c is the smallest non-negative number such that the equation

$$|Y|^{1/2} < (c+\varepsilon)\cdot |k|$$

has only a finite number of integer solutions (X,Y) for any choice of $\varepsilon > 0$.

According to [Dan] we have that $c > .0032$. Is $c \leq 1$?

The maximum number of integral solutions that Gebel, Petho, and Zimmer found for a single given equation (10) is 48 pairs $(\pm X, Y)$. Is the number of integral solutions on the Mordell equation uniformly bounded independent of k? My guess is that they are not.

Part II

§7. The passage from ABC to "cuspidal modular forms".

For this discussion, we will be assuming some knowledge of the theory of complex analytic functions of one variable. A **cuspidal modular form of weight two**, f(z), is a function of a complex variable z, convergent in the upper half-plane $z = x+iy$ for $y > 0$, having a Fourier expansion

$$(12)\quad f(z) = a_1 e^{2\pi iz} + a_2 e^{4\pi iz} + \ldots + a_n e^{2\pi inz} + \ldots$$

and such that for some choice of positive integer N (called a **level** for f) f(z) satisfies the transformation laws

$$f(Tz)\cdot d(Tz) = f(z)\cdot dz$$

for all linear fractional transformations $T(z) = az+b/cz+d$, with a,b,c,d integers, $ad-bc=1$, and c a multiple of N. For an introductory treatment of this subject, see 2.3 of [G], Ch. VII of [S 1], or [Mi]. To complete the definition of **modular form**, or of **cuspidal modular form**, one must also require a further technical condition which I won't describe fully except to say that for the complex analytic function f to be a **modular form** f must be holomorphic at all "cusps"; for it to be **cuspidal** f must be holomorphic and vanish at all "cusps"- the Fourier expansion of f displayed in (12) above guaranteeing this latter condition at the "cusp" $z=i\cdot\infty$.[9]

[9] For a definition and treatment of the notion of "cusps", see [Mi]; the **cusps** at level N are the points "at infinity" of the Riemann surface obtained by dividing the upper half-plane by the action of the group of linear transformations T (discussed in the paragraph above). One may develop the

For any fixed level N the vector space of modular forms of weight two is finite-dimensional. That is, for any fixed N we can select a finite set of modular forms $f_1,\ldots,f_s$ (of weight two and level N) such that any other modular form of weight two and level N is a linear combination of the f_j's. One has good numerical understanding of these modular forms and of their Fourier coefficients, at least for reasonably small level N. For example, there are no modular forms at all of weight two for level N=1. For any prime level N, there is at least one modular form $G_N(z)$ of weight two and level N (up to scalar multiplication) called the **Eisenstein series of weight two and level N**; its Fourier expansion is

$$(13) \quad G_N(z) = (N-1)/24 + \sum_{n=1}^{\infty} d_N(n)e^{2\pi inz}$$

where $d_N(n)$ is the sum of the positive divisors of n which are relatively prime to N. For each of the levels N=2,3,5 and 7, $G_N(z)$ is the only modular form (up to scalar multiplication) of weight two and of that level.

The Fourier coefficients of a modular form f(z), i.e., the a_n's occurring in the Fourier expansion (12), play an enormous role in the theory: on the one hand, these coefficients viewed as functions $n \mapsto a_n$ often have interesting arithmetic significance, and a particularly elementary example of this can be seen in (13); while on the other hand, various basic properties of, and

---

analytic function f(z) as a Laurent series in a local parameter in the neighborhood of each such cusp; the requirement that f be **holomorphic** is simply that this Laurent series be a power series; the requirement that f be **cuspidal** is that this power series have vanishing constant term. The requirement that f be cuspidal is equivalent to the growth condition $|f(x+iy)| \ll 1/y$ for all x, and $y > 0$.

interrelations between, modular forms are most directly seen in terms of these Fourier coefficients. The reader wishing to have more contact with this may turn to a number of excellent introductory and historical works listed in the bibliography. The central role that the Fourier coefficients $n \mapsto a_n$ themselves play in the theory of modular forms, and the recursive relations that, at times, bind these coefficients together is seen quite vividly in the theory of what are called **newforms.** To sketch this theory, let us say that two modular forms f and g of level N are "**almost equal**"[10] if $a_n(f) = a_n(g)$ for all integers n which are relatively prime to the level N, where $a_n(f)$ refers to the n-th Fourier coefficient of f, and $a_n(g)$ the same for g. A cuspform f(z) of level N (and weight 2) is defined to be a **newform** if f, viewed as a modular form of level N, is not "almost equal" to any modular form g of level strictly lower than N, and if the Fourier coefficients $a_n(f) = a_n$ satisfy these recursive relations:

(14) $a_1 = 1,$

$a_n \cdot a_m = a_{n \cdot m}$ if n and m are relatively prime,

$a_p \cdot a_{pm} = a_{p^2 \cdot m} + p \cdot a_m$ for all prime numbers p not dividing the level N,

$a_p \cdot a_{pm} = a_{p^2 \cdot m}$ for all prime numbers p dividing the level N.

---

[10] At first view, this may seem to be a somewhat disruptive thing to do to functions of a complex variable: to define an equivalence relation determined by consideration of *selected* subsets of their Fourier coefficients! This strategy grows on one, though, especially when motivated by the study of the action of Hecke operators on modular forms.

The systematic theory of **newforms** begins with work of Atkin and Lehner. It is an essential feature of this theory that the vector space of all cuspforms of level N (and weight 2) has a "chosen" basis comprised of modular forms all of which are "almost equal" to newforms of levels which are divisors of N. This chosen basis includes every newform of level N. The entire package of Fourier coefficients $\{a_n(f);\ n=1,2,3,\ldots\}$ of a newform f , and hence the newform itself, is reconstructable using the recursive relations listed in (14) if we are only given the Fourier coefficients $a_p(f)$ where p ranges through all prime numbers. It is, in fact, true that knowledge of the $a_p(f)$ for all but a finite number of primes p uniquely characterizes the newform f. In passing one might mention that the notion of newform is sometimes used in a slightly wider sense to incorporate certain noncuspidal modular forms as well (these are "Eisenstein series", an example of which is the modular form $G_2(z)$ in (13) above).

The recent work of Wiles, and Taylor-Wiles (and a more recent strengthening of these results due to F. Diamond; or see also [D-K]) showing that a large collection of elliptic curves defined by equations with integer coefficients

$$E: \quad Y^2 = X^3 + uX^2 + vX + w$$

are "modular" has been explained in a number of expository articles. There are many ways to express the fact that E is "modular" and here is one way: The elliptic curve E is modular if there exists a cuspidal modular newform of weight two and of some level N

$$f_E(z) = \sum_{n=1}^{\infty} a_n e^{2\pi i n z}$$

whose Fourier coefficients a_n are rational integers *and such that there is this miraculous link between* f_E *and* E:

The Link: For all but a finite number of prime numbers p, the number of solutions (X,Y) *in integers modulo* p of the cubic equation

$$Y^2 \equiv X^3 + uX^2 + vX + w \quad \cdot \bmod p$$

(*the same equation which defines the elliptic curve* E)

is given by the formula $p - a_p$ where a_p is the p-th Fourier coefficient of the newform f_E.

The p-th Fourier coefficients of the newform f_E, for all but a finite number of primes p, are determined (via the "link" above) by the elliptic curve E, and therefore the newform f_E satisfying this link to E is uniquely determined by E.

The Frey curve $E_{(\varepsilon)}$ of any ABC-solution (ε) is among the elliptic curves for which the Wiles, Taylor-Wiles, and Diamond results apply[11]. And so we can make the

[11] The more recent preprint [D-K] proves that all Frey curves are modular; this proof is based directly on the results of [W] and [T-W] and and is independent of [D]. It depends upon a calculation of the possible 2-parts of the level, using Tate's well known algorithm for reduction-types of elliptic curves.

passage:

$$(\varepsilon) \text{ ----> } E_{(\varepsilon)} \text{ -----> } f_{E_{(\varepsilon)}}$$

from ABC-solution to Frey curve and thence to the linked cuspidal newform of weight two, whose double-subscript notation $f_{E_{(\varepsilon)}}$ let us shorten to $f_{(\varepsilon)}$. A computation of the level N of the modular form $f_{(\varepsilon)}$ that we get from this passage gives that N is an explicit and relatively small power of 2, times the radical of A·B·C:

$$N = 2^{e} \cdot \text{rad}(A \cdot B \cdot C),$$

where e can be either -1,0,2 or 4.

But why is this transformation

$$(\varepsilon) \text{ -----> } f_{(\varepsilon)}$$

from ABC-solutions to modular forms so powerful a tool in the study of ABC-solutions?

The short answer to this comes in two parts, (A) and (B), below:

(A) Certain properties of the newform of level N

$$f_{(\varepsilon)} = \Sigma\, a_n e^{2\pi i n z}$$

associated, via the "link" above, to the Frey curve of an ABC-solution $(\varepsilon) = (A,B,C)$ are remarkably sensitive to the occurrence of perfect powers dividing A,B, or C.

What I mean by this will become clearer with the formulation of the "level-lowering principle", below.

First some standard notation: for q a prime number and M a nonzero integer, let us denote by $ord_q(M)$ the exponent of the highest power of q which divides M (e.g., $ord_2(24) = 3$ because 2^3 is the highest power of 2 dividing 24).

Fix r some prime number, and let q range through the odd prime number divisors of N for which

$$ord_q(A \cdot B \cdot C) \equiv 0 \text{ modulo } r.$$

(that is, q raised to a power which is a multiple of r is the highest power of q dividing $A \cdot B \cdot C$).

Let M be the product of all the above prime numbers q. So, M depends only upon r and N. For example, for the ABC-solution

$$2 + 3^{10} \cdot 109 + (-23^5) = 0,$$

and for r = 5, M is equal to $3 \cdot 23$.

Now a theory developed principally by Ribet which might be called the **level-lowering theory** (see [Ri 2-6], [Ca]) will guarantee, if $r > 3$ and $M > 1$, the existence of *another* cuspidal newform (call it g) of weight two and of level lower than that of f (the level of g will be $N/M \cdot 2^e$, where $e \geq 0$) such that Fourier coefficients of g are related to Fourier coefficents of $f_{(\varepsilon)}$ by congruences modulo r. We shall give a not-so-brief discussion of this congruence relation a bit later. But for the moment, let us just refer to the congruence relation by saying that the modular forms $f_{(\varepsilon)}$ and g are "**almost congruent modulo r**" so that we can formally display:

(15) "The Level-lowering principle for prime numbers r > 3" : *Let* (ε) *be an* ABC-*solution and*

$$f_{(\varepsilon)} = \sum_{n=1}^{\infty} a_n e^{2\pi i n z}$$

its associated newform. Let N *be its level, r a prime number* > 3, *and* M = M(N,r) *the integer defined as above. Then* $f_{(\varepsilon)}$ *is "almost congruent modulo* r " *to a cuspidal newform*

$$g(z) = \sum_{n=0}^{\infty} b_n e^{2\pi i n z}$$

of weight two and of level $N/M \cdot 2^e$, *for some non-negative integer* e.

To repeat: if we are given an ABC-solution (ε) where A,B, or C is divisible by a perfect power, besides getting the modular form $f_{(\varepsilon)}$ linked to its Frey curve we *also* get the prediction of the existence of some other newform (or forms) g of weight two, somehow connected to this ABC-solution (ε), but of comparatively *lower level* . The gain here comes from something we have already hinted, namely:

(B) We have a *very good* computational understanding of modular forms, cuspidal or merely holomorphic, of low level (and fixed weight).

It is now time to explain what it means for the coefficients of the two modular forms $f_{(\varepsilon)}$ and g to be **almost congruent modulo** r. The telegraphically brief

explanation would be just to say that we want $a_n \equiv b_n$ modulo r for all integers n that are relatively prime to the level N. But there is a technical glitch in this brief definition, for an important reason: although the level-lowering principle predicts the existence of a newform $g = \Sigma b_n e^{2\pi inz}$, it is *not necessarily the case that the Fourier coefficients* b_n *are rational integers*. If the b_n are all rational integers, then our telegraphically brief explanation above makes clear sense, and is, in fact, what we would mean by the assertion that $f_{(\varepsilon)}$ and g are "almost congruent modulo r". In general, the construction coming from the level-lowering principle does not always give newforms g all of whose Fourier coefficients are rational integers. But the set of all Fourier coefficients of any newform generates some number field, i.e., an extension of the field of rational numbers of finite degree. To say that $f_{(\varepsilon)}$ and g are **almost congruent modulo** r is to say that there is some maximal ideal m of the ring of algebraic integers $\mathcal{O}$ if the number field generated by the Fourier coefficients of g, where the quotient residue field $\mathcal{O}/m$ is of characteristic r (i.e., contains the field of rational integers modulo r) and

$$a_n \text{ (modulo r)} \quad \text{is congruent to} \quad b_n \text{ (modulo m)}$$

for all n relatively prime to N.

Celebrated Example. (For a fuller exposition of this example, see [Ma 1], [S 2], [G]). It is **(A)** and **(B)** together that wield the punch, and this is never more dramatically demonstrated than when we imagine ourselves presented with an ABC-solution which comes from a solution to the Fermat equation of prime exponent $r \geq 5$ $A=a^r$, $B=b^r$, $C=c^r$. For then we may take M to be the largest odd divisor of rad(a·b·c) and apply the level-lowering principle to find that our modular form $f_{(\varepsilon)}$ is congruent modulo r to a cuspidal newform g of weight

two and level an explicit power of 2: as it turns out, of level precisely equal to 2 if we label A,B, C to be such that A is congruent to -1 mod 4, and B is even. But as we have already mentioned, there is *no* cuspidal newform of weight two and that level. Hence the assumption that there is a solution to any Fermat curve of prime exponent ≥5 leads to a contradiction!

This kind of argument, which is a brand-new tool for finding all solutions to Diophantine equations, goes a good deal further, and it is a lot of fun to use it to analyze other equations, e.g., those of the type

$$(16) \qquad M \cdot X^n + Y^n + Z^n = 0$$

for coefficients M divisible only by a few small primes. For example, see ([S 2]; 4.3 Thm. 2; compare [G]) where such an analysis is given to show that (16) has no nontrivial solutions in integers X,Y,Z for prime exponents n ≥5, and for M any power of a prime p ≠n, for p taken from the set S= {3,5,7,11,13,17,19,23,29,53,59}. With more work, one can get this method to enlarge the set S of primes p for which (16) can be proven to have no nontrivial solutions, but the reader might note that at least some of the small primes *not* listed in S are excluded for good reason: e.g., p =2, p = 31. Even for some primes p for which solutions of (16) *actually exist* when M runs through powers of p, and n runs through exponents ≥ 5, this congenial method doesn't altogether abandon us: take the case of p=211, for which (16), taken with M = p, and n=5, has the solution

$$211 \cdot 1^5 + 2^5 + (-3)^5 = 0.$$

An amusing exercise in this method is to show that if ℓ = 211, or more generally, if ℓ is a prime number not of the form $2^a \pm 1$, i.e., neither a Mersenne prime nor a Fermat prime, then there is a bound n_ℓ so that no equation of the form (16) with M a power of ℓ, and with exponent $n > n_\ell$,

has a nontrivial solution in integers X,Y,Z. For a hint about how this works in the special case of $\ell = 211$, and in general, see the box labelled [*2] below. For the case of M equal to a power of 2, it has been conjectured by Dénes that (16) has only two nontrivial solutions for odd prime exponents n, i.e., M must be equal to 2, and

$$(X,Y,Z) = \pm(1, -1, -1).$$

See [Ri 7] for a discussion of this conjecture and for its verification in the case of prime exponents $n \equiv 1 \bmod 4$. For applications of this machinery (the Frey curve strategy, the modularity of such elliptic curves, and the level-lowering theory) to other Diophantine equations, see [Dar 1]and [D-G].

*******(put the box labelled [*2] here)**********

What about the "level-lowering principle" for the prime r = 3? The reason why we have excluded the case of r=3 is that although the "level-lowering principle" still works for r=3, it works with a slight change: Namely, the lower-level modular form g "almost congruent" modulo 3 to $f_{(\varepsilon)}$ which is guaranteed to exist by the "level-lowering principle" *need not be* a cuspform if r=3: it might be an Eisenstein series (of lower weight).

Let us relegate to the box labelled [*3] below the short technical discussion of these matters and mention that this contingency does happen, as we shall see in our first example below; in such cases the level can sometimes be lowered even further than predicted by the general "principle".

********(put the box labelled [*3] here)********

At this point I want to apologize for constantly talking

about "the" constructed newform g of lower level, in the discussion above. There is no claim to uniqueness of g: there may be many such g's that fit the bill.

To summarize our discussion so far, we may associate to any simple ABC-solution (ε) the following kind of dizzying constellation of modular forms:

$$
\begin{array}{lll}
 & \langle\text{---}\rangle & g_1 \\
(\varepsilon) \text{ ----} \rangle f_{(\varepsilon)} & \langle\text{---}\rangle & g_2 \\
 & & \vdots \\
 & & \vdots \\
 & & \vdots \\
 & \langle\text{---}\rangle & g_\nu
\end{array}
$$

where each g in this list is either a cuspidal newform of weight two which is "almost congruent" to $f_{(\varepsilon)}$ modulo some prime number r, and is of level $N/M \cdot 2^e$ (where M is the product of all odd primes q for which $\mathrm{ord}_q(A \cdot B \cdot C) \equiv 0 \bmod r$), or g is an Eisenstein series of lower level which is "almost congruent" to $f_{(\varepsilon)}$ modulo 3.

One may wonder: How much "almost congruence" of this sort can any single $f_{(\varepsilon)}$ have? Is there some perspicuous way of limiting, say, the prime numbers r occurring as moduli for "almost congruences" satisfied by $f_{(\varepsilon)}$? For an exposition of the surprising affirmative answer to this question, building upon the work of Shimura, Doi, Ohta, Hida, Ribet and others, the reader might consult the excellent article by Zagier [Z] (See also Prop. 4.1 of [Wi]); for the briefest hints about this see the box labelled [*4] below.

*******(put the box labelled [*4] here)**********

In [Dar 2], Darmon has formulated some conjectures (Conj. 4.4, Conj. 4.5 of loc. cit.; these are strengthenings of an earlier Conjecture of Frey) which have implications about the extent to which "almost congruences" can occur between two newforms of weight two with *rational integral* coefficients. In particular, a consequence of Frey-Darmon's Conjecture 4.4 is the following

Conjecture (Frey, Darmon): *There is a constant* $B < \infty$ *such that if* f *and* g *are any two distinct newforms of weight two, each with rational integral coefficients, which are "almost congruent" to each other modulo* p, *for* p *a prime number, then*

$$p \leq B.$$

For any prime number $p \geq 7$, one can show[12] that there are only a finite number of distinct newforms g of level two, with rational integral coefficients, which are in the same "almost congruence class" modulo p.

Question: For a given p, how many pairs of distinct newforms f,g are there, with rational integral coefficients, which are "almost congruent" mod p?

Quite a number of such pairs are known for p =7, and I am thankful to John Cremona for making calculations which suggest that[13] for newforms of level ≤ 5000 there

[12] applying Faltings Theorem to an appropriately twisted model of the modular curve of full level p structure

[13] By "suggest that" I mean to indicate that Cremona lists a pair f,g as being possibly "almost congruent" modulo p if the first 1000 of their Fourier coefficients for exponents relatively prime to the levels involved are

are seventeen such pairs for p = 11, two such pairs for p = 13, and no such pairs for any prime number p greater than 13. None of the above pairs discovered by Cremona are newforms attached to Frey curves.

One can think of the above question for a given prime p as being a problem about rational points on the Hilbert modular surface (call it H(p)) attached to the order of index p in $\mathbb{Z}\times\mathbb{Z}$. This surface has a natural $\mathbb{Q}$-structure and for this $\mathbb{Q}$-structure, any pair of elliptic curves (E,E') together with an isomorphism between their $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$-modules of p-torsion points determines a $\mathbb{Q}$-rational point on H(p). If N is a positive integer relatively prime to p, the curve $X_0(N)$ (whose points classify N-isogenies $E\to E'$ of elliptic curves) maps naturally to H(p) : the map sends the point of $X_0(N)$ corresponding to the N-isogeny $E\to E'$ to the point in H(p) corresponding to the pair (E,E'). Assuming the Shimura-Taniyama-Weil Conjecture, any noncuspidal rational point of H(p) which does not lie on any of the images of these curves $X_0(N)$ gives rise to a pair (f,g) as described in the Question above. Elkies has found a rational parametrization of H(7), which explains the proponderance of such couples (f,g) for p=7. In the recent preprint of Kani and Schanz [K-S] it is shown that the Hilbert modular surface H(7) is rational, H(11) is a (blown-up) elliptic surface, and H(p) is a surface of general type for $p \geq 13$. Are there primes $p \geq 13$ for which one can find a curve of genus < 2 lying on H(p) which is not the image of one of the modular curves $X_0(N)$? In view of some well known conjectures of Lang, this is directly relevant to the study of the rational points on these surfaces. The arithmetic, and the algebraic geometry of these modular surfaces seem very much worth exploring.

---

congruent modulo p. It would be an interesting exercise to rigorously prove that each of Cremona's pairs are indeed "almost congruent" modulo p.

Examples of "almost congruence": The point of the "Celebrated Example" discussed above, was that the "level-lowering principle" ruled out the existence of ABC-solutions of a certain type. To get a more intimate sense, though, of the nature of this "level-lowering principle" let us exhibit a few ABC-solutions (i.e., ones that actually do exist!) to see how "level-lowering" plays out with them. See also §3 of [Dar 2] for more examples.

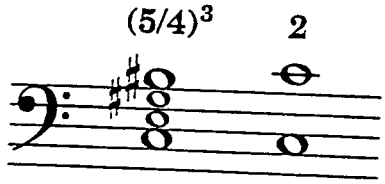
I. r=3. The lower level newform g needn't be cuspidal

Consider the ABC-solution

(ε_1) $\quad (-3) + (-5^3) + 2^7 = 0.$

whose associated Frey curve $E_{(\varepsilon_1)}$ is the one labelled 30A (F) in the extremely useful book of tables of Cremona [C].

I am thankful to Elkies for suggesting this ABC-solution to me and also for suggesting his (useful??) mnemonic for it:



This is meant to illustrate how "very close" $(5/4)^3$ is to 2. Namely, 5:4 is the ratio of frequencies of upper to lower notes in a major third; and the ratio of top to bottom notes in three major thirds, one on top of the other, i.e.

$(5/4)^3$:1 , is sufficently close to that of the octave, i.e. 2:1, that the two notes $B^{\#}$ and C are merged as one on well-tempered instruments-- this closeness being reflected, as well, by the rather respectable power P = 1.42656... of the ABC-solution (ε_1) above. This ABC-solution is the 30-th entry in the table of [B-B] .

The corresponding newform $f_{(\varepsilon_1)}$ is of level 30, and writing q for $e^{2\pi iz}$ in its Fourier expansion, $f_{(\varepsilon_1)}(z) = \Sigma\, a_n q^n$, its first few terms are

$$f_{(\varepsilon_1)}(z) = q + q^2 - q^3 + q^4 + q^5 - q^6 - 4q^7 + q^8 + q^9 + q^{10} - q^{12} \ldots \ .$$

The Fourier coefficients a_n, for n relatively prime to 30, satisfy the congruence

(17) $\qquad a_n \equiv d_n \pmod 3$

where d_n is the sum of all positive divisors of n.[14] Now there are no cuspidal newforms of level 6 or of level 3 of weight 2, but there are Eisenstein series, and the congruence (17) tells us that $f_{(\varepsilon_1)}$ is "almost congruent" modulo 3 to each of the Eisenstein series G_3 or G_2 defined in (13). Our "constellation", in this case looks like:

---

[14] This Frey curve admits a rational 3-isogeny .

ABC-sol'n	associated newform		lower level (noncuspidal) modular form
(ε_1) --->	$f_{(\varepsilon_1)}$	<------->	G_3 level= 3
	N= 30	<------->	G_2 level= 2
		almost congruent mod 3 .	

II. r=3. **The lower level newform g might be** cuspidal. Consider the ABC-solution

(ε_2) $\quad 3^3 + 2^4 + (-43) = 0$

whose newform $f_{(\varepsilon_2)}$ is of level N = 129 =3·43 and corresponds to the elliptic curve 129B in [C]. The level-lowering principle guarantees a modular form g of weight two of level 43 which is "almost congruent" to $f_{(\varepsilon_2)}$ modulo 3. In this example, however, the dimension of the vector space of cuspforms (of weight two) of level 43 is 3. There are three cuspidal newforms of level 43, weight 2, forming a basis of the space of all cuspforms of that level and weight. Precisely one of these newforms (call it g) has its Fourier coefficients rational integers, while the other two have Fourier coefficients in the quadratic field obtained by adjoining $\sqrt{2}$ to the field of rational numbers (these two newforms being "conjugate"). The modular form $f_{(\varepsilon_2)}$ is "almost congruent" modulo 3 to g and is not "almost congruent" modulo r, for r an odd prime, to (either of) the other two newforms or to an Eisenstein series, so the "constellation" looks like:

ABC-sol'n	associated newform		lower level newform
(ε_2) --->	$f_{(\varepsilon_2)}$	<------->	g
	N= 129	almost congruent mod 3	N/M = 43

III. r=5.

Now consider the ABC-solution

(ε_3) $\quad 13 + 3^5 + (-2^8) = 0$

whose power P is 1.2727... and whose Frey curve $E_{(\varepsilon_3)}$ is the one labelled 78A2 in [C]. By the level-lowering principle, there should exist a newform of level $26 = 2 \cdot 13$ and of weight two whose n-th Fourier coefficients (for n relatively prime to 78) are congruent to those of the newform $f_{(\varepsilon_3)}$ modulo 5. There are, in fact, *two* cuspidal newforms of weight two and level 26, both of whose Fourier coefficients are rational integers. Call these cuspforms g and h, corresponding (in the sense, e.g., of "The Link" above) to the modular elliptic curves labelled 26A and 26B, respectively, in Cremona's tables. One checks that it is the cuspform g which is "almost congruent" modulo 5 to $f_{(\varepsilon_3)}$.

Our "constellation" in this case looks like:

ABC-sol'n	associated newform		lower level newform
(ε_3) --->	$f_{(\varepsilon_3)}$	<------->	g
	N= 78	almost congruent mod 5	N/M = 26

IV. r= 3 and 5, and an example where a lower level newform has coefficients generating a quadratic number field.

Try the ABC-solution:

(ε_4) $\qquad (-5^3) + 3^5 + (-2^4 \cdot 23) = 0.$

Here $f_{(\varepsilon_4)}$ is of level $3 \cdot 5 \cdot 23$ and the associated Frey curve is the one labelled 345C in [C]. By the level-lowering principle, $f_{(\varepsilon_4)}$ is "almost congruent" modulo 5 to some cuspidal newform (of weight two) of level 115. There are seven newforms of that level, one whose Fourier coefficients are rational integers, two (conjugate) newforms g_1, $\bar{g}_1$, with Fourier coefficients in the field obtained by adjoining $\sqrt{5}$ to the field of rational numbers, and four more (mutually conjugate) newforms with Fourier coefficients generating a field of degree 4. One checks that $f_{(\varepsilon_4)}$ is "almost congruent" modulo 5 to both g_1 and to $\bar{g}_1$, but satisfies no further "almost congruences"[15] modulo 5. The level-lowering principle

[15] I thank Siman Wong for his program which helped with this check

also tells us that $f_{(\varepsilon_4)}$ is "almost congruent" modulo 3 to a modular form of level 69; one checks that this modular form must be a cuspidal newform, and there are three to choose from at this level, one with rational integer coefficients, call it g_2, "linked" to the elliptic curve labelled 69A in [C], and two other (conjugate) newforms each with Fourier coefficients in the quadratic field obtained by adjoining $\sqrt{5}$ to the field of rational numbers. Our constellation, then, looks like:

ABC-sol'n	associated newform		lower level newforms
(ε_4) --->	$f_{(\varepsilon_4)}$	<------->	g_1 , $\bar{g}_1$
	N= 345	almost congruent mod 5	N/M = 115
		<------->	g_2
		almost congruent mod 3	N/M = 69

§8. Passage from ABC-solutions to algebraic points on a Fermat curve. As if the two transformations,

$(\varepsilon) \dashrightarrow E_{(\varepsilon)}$ and $(\varepsilon) \dashrightarrow f_{(\varepsilon)}$

and the attendant constellation of modular forms g of "lower level" variously congruent to $f_{(\varepsilon)}$ already described were not enough mathematical structure to impose on a mere solution to the ABC-equation, Vojta and others have proposed yet another kind of transformation, with quite a different spirit to it, which links ABC-solutions to yet

another branch of arithmetic and algebraic geometry. As with the Frey construction, this construction is elementary to describe but involves a certain amount of choice, and one might very well be struck by its curiously ad hoc appearance. Nevertheless, it has the virtue of thrusting ABC-solutions into a branch of arithmetic and algebraic geometry where a different range of techniques might (one day!) be brought to bear on the study of the ABC-conjecture.

To the ABC-solution

$$(\varepsilon) \qquad A + B + C = 0,$$

Vojta associates the collection of algebraic points (α,β,γ) on the Fermat curve

$$F : X^4 + Y^4 + Z^4 = 0$$

where (the reader may have guessed that we want)

$$\alpha^4 = A, \quad \beta^4 = B, \quad \gamma^4 = C.$$

One could have equally well replaced the '"curve F of exponent 4" with the Fermat curve of exponent n for some fixed $n \geq 4$ and then give a corresponding definition for the collection of algebraic points. The motivation for performing this curious operation

(18) (ε) ---> the set of such algebraic points

is that our curve F and these algebraic points are objects of study in a relatively new subject with some new techniques at its disposal. The subject, initiated by the Russian mathematician Arakelov in honor of whom it is called "Arakelov Theory" (and its higher dimensional analogue which is usually called "Arithmetic Algebraic

Geometry") has been developed and is currently being refined by a number of mathematicians, including Szpiro, Soulé, Gillet, Bismut, Vojta, Faltings, Bost, Zhang, Burnol, and Kim.

"Arithmetic Algebraic Geometry" is a synthesis of arithmetic and of classical algebraic geometry: it captures Minkowski's "geometry of numbers", the classical theory of algebraic surfaces, the analytic theory of Hermitian line bundles on Riemann surfaces, and the arithmetic theory of algebraic curves and their rational points, all in one unified setting. It provides a geometric format for some of the standard constructions in transcendental number theory. It has deep ties with Nevanlinna Theory.

In Arakelov Theory, the Fermat curve F (and, in general, any algebraic curve of genus ≥ 2) is given a suitable structure so as to allow it to be treated as somewhat analogous to a "surface S of general type" in the classical theory of algebraic surfaces, an algebraic point P on F being analogous to a curve C on a classical surface S. The "size" or "height" of algebraic points P is analogous, in the classical picture, to the degree of the canonical bundle of S restricted to the curve C. In 1986, Parshin, pursuing this analogy, made some conjectures in Arakelov Theory which are analogous to known classical inequalities in the theory of algebraic surfaces-- these conjectures of Parshin (still unproven) having strong consequences concerning the size of algebraic points. In this vein, the ABC-Conjecture becomes a piece of a larger philosophy due primarily to Vojta. The interested reader can consult the appendix to [L] written by Vojta, for an account of these conjectures, and for the surprising proof that if one applies Parshin's conjecture to the collection of algebraic points produced by the rule (18) above, one would get the ABC-conjecture as a consequence.

What lies ahead for ABC? For Arithmetic Algebraic Geometry? The drama of Mathematics being such that we

usually have no idea what shape our subject will take in the future, this is probably the right point to end an article for a volume entitled New Directions in Mathematics... except for two appendices.

Part III (Appendices)

Appendix A: A hint about how ("ABC" implies "Mordell").

We will write down a neat inequality, due to Elkies [E], which is the key to the connection between "ABC" and "Mordell". The shape of the underlying argument which makes use of this inequality is in the tradition of the well-known constructions that connect the occurrence of integral points on certain algebraic curves to rational approximations of certain algebraic numbers.

Let us first give more concise "packaging" to ABC-solutions: Given an ABC-solution (A,B,C), let r=r(A,B,C) be the rational number -A/B. So r is a rational number distinct from 0 and 1 (for if not, then A,B, or C would have to be 0, which is not allowed) and, since A,B,C have no common divisors, we can reconstruct A,B, and C from r. Therefore, the set of ABC-solutions is in one:one correspondence with the set $\mathbb{Q}-\{0,1\}$, where $\mathbb{Q}$ is the field of rational numbers, or (what amounts to the same thing) with the set of rational points different from 0,1, or ∞ on the projective line, i.e., the set $\mathbf{P}^1(\mathbb{Q}) - \{0,1,\infty\}$. Our "power" function P on ABC-solutions may now be viewed as a curious function,

$$P : \mathbf{P}^1(\mathbb{Q}) - \{0,1,\infty\} \rightarrow \text{positive reals.}$$

$$r = r(A,B,C) \mapsto \log(\max(|A|,|B|,|C|))/\log(\mathrm{rad}(A\cdot B\cdot C)).$$

Let, now, X be a smooth (projective) algebraic curve defined over $\mathbb{Q}$. We use the usual notation X(K) for its set

of K-rational points for K any field containing $\mathbb{Q}$; e.g., if $\bar{\mathbb{Q}}$ is an algebraic closure of $\mathbb{Q}$, then $X(\bar{\mathbb{Q}})$ is the set of algebraic points on X.

Let f be any nonconstant rational function from X to P^1,

$$f: X \rightarrow P^1,$$

and let d = degree(f) and m = the number of points of $X(\bar{\mathbb{Q}})$ whose image under f is 0,1, or ∞ (the actual number, without taking account of multiplicities).

If α is any rational point of X which does not map to 0,1, or ∞ under f, then $r = f(\alpha)$ corresponds to an ABC-solution (A,B,C), and therefore the power function P is defined on $f(\alpha)$. Using a fairly direct analysis, Elkies shows the following inequality:

$$P(f(\alpha)) > d/m + \mathcal{E}(\alpha) \tag{19}$$

where $\mathcal{E}(\alpha)$ is an "error term" bounded as follows:

$$\mathcal{E}(\alpha) \leq C/(\log(\max(|A|,|B|,|C|)))^{1/2} \tag{20}$$

where C is an effective constant dependent upon X, and f, but not upon α.

Suppose we are given an infinitude of rational points α in $X(\mathbb{Q})$. Ignoring those (at most m of them) that map to 0,1, or ∞, we may suppose that the inequality (19) holds for $P(f(\alpha))$, and since there are only a finite number of ABC-solutions (or triples of integers, for that matter) with bounded $\log(\max(|A|,|B|,|C|)$, by (18) we have

$$\lim P(f(\alpha)) \geq d/m \tag{21}$$

the limit taken over any infinite sequence of α's in $X(\mathbb{Q})$.

At this point, let us assume that our curve X is of genus

≥ 2, and make use of a construction of Belyi [Be]. Belyi provides us with a (nonconstant) rational function f, defined over $\mathbb{Q}$, on any smooth projective curve X, of genus ≥ 2 defined over $\mathbb{Q}$, which has the property that the number of distinct points of $X(\bar{\mathbb{Q}})$ which map to the set $\{0,1,\infty\}$ is strictly less than the degree of f. That is,

(22) $\quad d/m > 1.$

The existence of such an f is, in fact, equivalent to the statement that the smooth projective curve X is of genus ≥ 2.

Now if X is of genus ≥ 2, and supposing that we are supplied with a nonconstant rational function f satisying (22) as guaranteed by Belyi's Theorem, we get (under the assumption of an infinity of elements in $X(\mathbb{Q})$) a straight violation of the ABC-Conjecture from (21). Therefore ABC implies "Mordell", and also, an appropriately effective version of ABC will translate to an effective version of "Mordell". If, as Elkies does in [E], we apply this same construction to an elliptic curve X having an infinity of rational points, and a function f on X with $m = d$ (in this case we cannot find an f with $m < d$), we would get an infinity of ABC-solutions (ε) such that lim.inf. of the power function $P(\varepsilon)$ is ≥ 1 (and therefore, assuming ABC, the limit of $P(\varepsilon)$ is equal to 1).

Appendix B. Consecutive perfect powers

Shorey and Tijdeman's book [S-T] gives an excellent discussion of the proof of Tijdeman's Theorem. Very briefly, one reduces the question immediately to the case of an equation of the form

(23) $\quad X^p - Y^q = \varepsilon$

where $\varepsilon = \pm 1$, p, q distinct prime numbers, $p > q$, both reasonably

large, and we assume that (x,y) is a solution with x and y relatively prime, and (of necessity) $x < y$. The first step in the proof, as in the classical proofs of Fermat's Last Theorem for regular prime exponents, is a "descent" of sorts. That is, write

(24) $x^p = y^q + \varepsilon$ and (25) $y^q = x^p - \varepsilon$.

Now, the right-hand-side of (24) factors as $(y+\varepsilon)$ times the integer

$$(y^q + \varepsilon)/(y + \varepsilon)$$

which one easily sees is either relatively prime to $(y+\varepsilon)$ or, at worst, shares a common divisor of q with $(y+\varepsilon)$. It follows that $y+\varepsilon$ is a perfect p-th power, except for the possible factor of this common divisor; in equations, if δ denotes an integer which is either 0 or 1, then

(26) $$y + \varepsilon = q^{\delta} s^p.$$

The same remarks for equation (25) give us

(27) $$x - \varepsilon = p^{\gamma} r^q$$

for γ an integer which is either 0 or 1. As Cameron Stewart pointed out to me, this step, yielding equations (26) and (27), is the major obstruction to extending Tijdeman's proof to more general equations (e.g., to equations of the form $X^m - Y^n = k$ for $k = 2,3,\ldots$). Also critical for the estimates to take place in the Theorem is the fact (easily worked out from these equations) that r and s are roughly of the same size. Putting these equations together, we have integers (r,s) satisfying:

(28) $$(p^{\gamma} r^q + \varepsilon)^p - (q^{\delta} s^p - \varepsilon)^q = \varepsilon.$$

The rest of the proof consists of making two applications of Baker's lower bound (e.g., Theorem B.1 of [S-T]) the first to show

(29) $q \ll (\log p)^4$,

and the second to show

(30) $p \ll (\log q)^7$;

where " A << B" means that A is less than an effectively computable constant times B.

The bounds (29), (30) together give us that we need only consider an (effectively computable) finite number of equations of the form (23); for each one of these equations, Baker's method provides an effectively computable upper bound to the number of its solutions, therby giving Tijdeman's Theorem.

To get the first bound (29), one (assumes, first, that q is quite large with respect to p, and then) estimates the quantity $\Gamma_1 = p^{\gamma} r^q / q^{\delta} s^p$ as being close to 1, in the sense that the absolute value of its logarithm is $\leq 12p^3 r^{-q}$. But this quantity Γ_1 is not equal to 1, and a direct application of Baker's theorem to its logarithm, written as the linear form

$$\log \Gamma_1 = p\gamma \cdot \log p - q\delta \cdot \log q + pq \cdot \log(r/s)$$

in log p, log q, and log (r/s), gives that its absolute value is greater than $r^{-c.\log(p)^4}$. Comparing these two estimates on $|\log \Gamma_1|$ gives (29). A similar argument with the quantity $\Gamma_2 = (p^{\gamma} r^{q+\varepsilon})^p / (q^{\delta} s^p)^q$ applying Baker's theorem to $\log \Gamma_2$, viewed as the linear form

$$\log \Gamma_2 = -q\delta \cdot \log q + p \cdot \log ((p^{\gamma} r^{q+\varepsilon})/s^q)$$

in log q and $\log ((p^{\gamma} r^{q+\varepsilon})/s^q)$ gives (30); see [S-T].

Bibliography

[Ba 1] Baker, A.: Effective methods in Diophantine problems, I, II, Proc. Symposia Pure Math. **20**, A.M.S. 1971 (pp. 195-205); **24** (pp. 1-7)

[Ba 2] Baker, A.: A sharpening of the bounds for linear forms in logarithms, I,II,III, Acta Arith. **21** (1972) 117-129; **24** (1973) 33-36; **27** (1975) 247-252

[Ba 3] Baker, A.: Transcendental Number Theory, Cambridge Univ. Press (1975)

[Ba 4] Baker, A.: Review of Catalan's Conjecture by Paulo Ribenboim, Bull. A.M.S. **32** (1995) 110-112

[Be] Belyi, G.: On the Galois extensions of the maximal cyclotomic field, in Russan, Izv. Akad. Nauk SSSR **43** (1979) 267-276

[B-B] Browkin, J., Brzeziński, J.: Some remarks on abc - conjecture, Preprint.

[C-H-M 1] Caporaso, L., Harris, J., Mazur, B.: Uniformity of rational points. To appear in the Journal of the A.M.S.

[C-H-M 2] Caporaso, L., Harris, J., Mazur, B.: How many rational points can a curve have? pp. 13-31 in The Moduli Space of Curves (Eds: R. Dijkgraaf, C. Faber, G. van der Geer) Progress in Mathematics **129**, Birkhäuser (1995)

[Ca] Carayol, H.: Sur les représentations galoisiennes modulo ℓ attachées aux formes modulaires, Duke Math J. **59** (1989) 785-801

[Co] Cox, D.: Introduction to Fermat's Last Theorem, American Math. Monthly **101** (1994) 3-14.

[C] Cremona, J: Algorithms for Modular Elliptic Curves, Camb. Univ.

Press (1992)

[Dan] Danilov, L.: The diophantine equation $y^2-x^3=k$ and a conjecture of M. Hall (Russian) Mat. Zametki **32** (1982) 273-275. Corr. **36** (1984) 457-458. Engl. Trans.: Math. Notes **32** pp. 617-618; **36** p. 726.

[Dar 1] Darmon, H.: The equations $x^n+y^n=z^2$ and $x^n+y^n=z^3$, Int. Math. Res. Notices **10** (1993), 263--274.

[Dar 2] Darmon, H.: Serre's Conjectures, Preprint (Rapports CICMA reports Concordia Laval McGill) 1994. To appear in Elliptic Curves, Galois Representations and Modular Forms, CMS Conf. Proc., AMS Publ., Providence.

[D-D-T] Darmon, H., Diamond, F., Taylor, R.: Fermat's Last Theorem, pp. 1- 107 in Current Developments in Mathematics, 1995 International Press (1995)

[D-G] Darmon, H., Granville, A.: On the equations $z^m=F(x,y)$ and $Ax^p + By^q = Cz^r$. To appear in the Bulletin of the London Math Soc.

[D] Diamond, F.: On deformation rings and Hecke rings. Preprint, Cambridge Univ. Nov. 1994.

[D-K] Diamond, F., Kramer, K.: Modularity of a family of elliptic curves. Preprint 1995.

[Ed] Edwards, H.M.: Fermat's Last Theorem: A genetic introduction to algebraic number theory, Graduate Texts in Math. **50**, Springer-Verlag, 1977.

[E] Elkies, N.: ABC implies Mordell, Int. Math. Research Notices **7** (1991) 99-109

[Fi] Fibonacci (Leondardo Pisano) The Book of Squares (annotated Engl. translation by L.E. Sigler) Academic Press (1987)

[Fr] Frey, G.: links between solutions of A-B=C and elliptic curves,

pp. 31-62 in Number Theory, Ulm 1987, Proceedings, Lecture Notes in Math. 1380 Springer-Verlag, 1989.

[G-P-Z] Gebel, J., Pethö, A., Zimmer, H.: On Mordell's Equation, preprint 1995

[G-M-O-S] Glass, A., Meronk, D., Okada, T., Steiner, P.: A small contribution to Catalan's equation, Journal of Number Theory 47 (1994) 131-137

[G] Gouvêa, F.: "A marvelous proof", American Math. Monthly 101 (1994) 203-222

[H-R] Hayes, B., Ribet, K: Fermat's Last Theorem and modern arithmetic, American Scientist **82** (1994) 144-156

[He] Hellegouarch, Y.: Points d'ordre $2p^h$ sur les courbes elliptiques, Acta Arith. **26** (1975) 253-263

[K-S] Kani, E., Schanz, W.: Diagonal quotient surfaces, preprint 1995.

[K-K] Kulesz, L., Keller, W.: Courbes algébriques de genre 2 et 3 possédant de nombreux points rationnels, preprint (1995)

[L] Lang, S.: Introduction to Arakelov Theory , Springer-Verlag (1988)

[L-M-N] Laurent, M., Mignotte, M., Nesterenko, J..: Formes lineaires en deux logarithmes et déterminants d'interpolations. To appear in Journal of Number Theory.

[Mas] Masser, D.: Open problems, in: Proc. Symp. Analytic Number Theory (W.W.L. Chen, ed.) London: Imperial College (1985)

[Ma 1] Mazur, B.: Number Theory as gadfly, American Math. Monthly **98** (1991) 593-610

[Ma 2] Mazur, B.: Arithmetic of Curves, Bull. A.M.S. ### (14) (1986) 207-259

[Wi] Wiles, A.: Modular elliptic curves and Fermat's Last Theorem, Annals of Math **141** (1995) 443-551.

[Z] Zagier, D.: Modular parametrization of elliptic curves, Canad. Math. Bull. **28** (1985) 372-384

[Z-K] Zagier, D., Kramarz, G.: Numerical investigations related to the L-series of certain elliptic curves, J. Indian Math. Soc. **52** (1987) 51-60