THE UNITY AND BREADTH OF MATHEMATICS—FROM DIOPHANTUS TO TODAY

BARRY MAZUR

Background notes for the 2018 Paul Bernays Lectures:

(i) **What is it that unifies Mathematics?**
Tuesday, September 11, 2018, 17.00 h

(ii) **New issues, and expectations, in the study of rational points**
Wednesday, September 12, 2018, 14.15 h

(iii) **Diophantine stability and the vanishing of L-functions at the central points**
Wednesday, September 12, 2018, 16.30 h

Contents

Part 1. What is it that unifies Mathematics?

ABSTRACT. Is mathematics unified? Does it matter? And are there useful ways to think about these questions?

Geometry as Algebra and *Algebra as Geometry*—these metaphors have been with us since ancient times and and the sheer wonder has never faded. Nowadays, mathematicians are attempting to join great mathematical fields—each with their own distinct brand of intuition—into grand syntheses enjoying intuitive power not matched by any single field alone. The Langlands Program is one on-going example of this. Model Theory offers a majestic and unified overview of the fundamental syntax we use. Powerful computing is now at the heart of many theoretical pursuits. There is a resurgence of interest in what is called *arithmetic statistics* (which I'll describe in later lectures). And the range of our applications is expanding. I'll give an overview of these issues and the questions they raise.

I imagine that if a biologist were asked for a single word that would appropriately point to the essence and substance of biology, the word would be *Life*. It stands for the essential unity of that subject despite the enormous range of different interests of biologists—from proteins to the behavior of elephants to medical applications.

Or for Economics, a single word that captures the unity of the enterprise might be *Exchange*.

Is there an analogous 'unifying anchor' that signals the vast range of mathematical sensibilities, accomplishments, emerging intuitions, truths, and applications, of mathematics? Besides, of course, the word *mathematics* itself, being given a somewhat circular definition.

One of the delights of Mathematics is that—despite the vastness of its range, the depth of its ideas, and the multitude of temperaments that engage in it—the subject is all bound together by an illuminating, and often surprising, fabric of connections. One is constantly made aware of new connections that increase our understanding, and capability for further understanding[1].

Distinct mathematical fields—each with their own distinct brand of intuition—combine to form grand syntheses enjoying intuitive power not matched by any single field alone: consider Combinatorial group Theory, Algebraic Geometry, *Arithmetic* Algebraic Geometry, the yoke

[1]Just to cite one very recent example: in the solution of the sphere packing problem in dimension 8 by Maryna Viazovska, a certain mock-modular form (such objects being the invention of Ramanujan) makes a very surprising—to me—appearance, and plays a crucial role.

between Algebraic Geometry and Symplectic Geometry offered by mirror symmetry, etc.

There is simply an over-arching unity to our subject. In these *Paul Bernays Lectures*—and contemplating the unifying spirit of Bernays-Hilbert's grand *Grundlagen* of Mathematics—I would like to examine just a few of the many facets of this unity (despite the fact that 'many-faceted unity' already sounds like a contradiction in terms).

We'll review how mathematics is unified by

- common and abiding interests ranging through centuries, through millennia,
- common language, definitions, and modes of expression, and common, coherent, ways of asking questions,
- common foundations, common 'substrate,'
- analogies,
- parallel or surprisingly compatible structures,
- and, perhaps, common goal.

1. Unity through time: from Diophantus to today

Modern mathematicians can sometimes find themselves in close conversation with ancients. For example, take *Problem VI.17* of (the third Arabic book of) Diophantus[2], this dating from the 3-rd century AD:

> Find three squares which when added give a square, and such that the first one is the *side* (i.e., the square-root) of the second, and the second is the *side* of the third.

If we interpret this as the quest for positive rational numbers that have the above properties, Diophantus himself offers a solution, along with a hint about how he arrived at it:

$$(\frac{1}{2})^2 \quad + \quad (\frac{1}{2})^4 \quad + \quad (\frac{1}{2})^8 \quad = \quad \frac{64+16+1}{256} \quad = \quad (\frac{9}{16})^2$$

[2] For a discussion of the works of Diophantus and its reception, see Norbert Schappacher's marvelous *Diophantus of Alexandria: a Text and its History* http://irma.math.unistra.fr/~schappa/NSch/Publications_files/1998cBis_Dioph.pdf.

The hint is—in effect—to notice that the square of $a + \frac{1}{2}$ is

$$a^2 + a + (\frac{1}{2})^2,$$

so if you take a to be $(\frac{1}{2})^4$, you win. Of course, what Diophantine *won* was this single solution. A more modern turn on such problems is often to quantify goals more precisely. Eg., *find* **all** *solutions (of the above problem) in positive rational numbers.*

This was achieved a mere 17 centuries later. That is: *Diophantus's solution is the only solution*[3]. The issue here is to find the rational points on the curve:

$$C : y^2 = x^6 + x^2 + 1, \tag{1.1}$$

solvable thanks (only) to relatively recent extensions of what is known as *Chabauty's method* for dealing with such problems: the only positive rational solution is given, as discussed above, by $x = \frac{1}{2}$. I'll be returning to this with some more discussion later.

That such cross-century conversations can be fruitful, and coherent, attests to some stability in what one might call our shared sensibilities regarding our subject, and also in—if not our precise language, at least in—the general way we allude to the substances that interest us. To help with this, Mathematics makes use of—I won't call them 'languages' since they aren't exactly languages, but rather—*modes of expression, or modes of operation* that bring things together.

2. Modes of expression, modes of operation

Ancient organizational schemes of logic, such as the *Organon* of Aristotle, have been vastly influential and have been—even if largely implicit—the armature of the way in which we formulate assertions, ask questions, and reach conclusions in mathematics as in everything else. Aristotle begins his discussion in the *Prior Analytics* by pinning down "deduction," (*syllogism*) as being

[3] See Joseph L. Wetherell's 1998 Berkeley thesis: "Bounding the number of rational points on certain curves of high rank, and Schappacher's discussion of this (page 24 of loc.cit.).

> discourse in which, certain things being stated, something other than what is stated follows of necessity from their being so.

Defining things:

Since *definition*, defined by Aristotle as: *an account which signifies what it is* **to be** *for something*[4] plays such a vital role in mathematics, the notion deserves some thought. Mathematics seems to require as strict lack-of-ambiguity in its assertions as possible, and therefore maximal clarity in its definitions. But perhaps—since ambiguity is sometimes unavoidable—it is better to say that any ambiguity should be unambiguously labeled as such.

The nature, and role, of *definition* in mathematical usage has evolved in remarkable ways. Consider the first few definitions in Book I of Euclid's *Elements:*[5]

(i) A **point** is that which has no part.
(ii) A **line** is breadthless length.
(iii) A **straight line** is a line which lies evenly with the points on itself.

and their counterparts in Hilbert's rewriting of Euclid's *Elements*, which begins with:

> Let us consider three distinct systems of things. The things composing the first system, we will call **points** and designate them by the letters A, B, C,. . . ; those of the second, we will call **straight lines** and designate them by the letters a, b, c,... The points are called the **elements of linear geometry**; the points and straight lines, **the elements of plane geometry**...

[4] a puzzling definition: *logos ho to ti ên einai sêmainei*

[5] These 'Elements' have quite an impressive spread, starting with the proclamation that a point is characterized by the property of 'having no part,' and ending with its last three books, deep into the geometry of solids, their volumes, and the five Platonic solids. It is tempting to interpret this choice of ending for the *Elements* as something of a response to the curious interchange between Socrates and Glaucon in Plato's *Republic* (528a-d) where the issue was whether Solid Geometry should precede Astronomy, and whether the mathematicians had messed things up.

It also would be great to know exactly how—in contrast—the *Elements* of Hippocrates of Chios ended. (It was written over a century before Euclid's *Elements* but, unfortunately, has been lost.)

One might call Euclid's and Hilbert's formulations **primordial definitions** since they spring ab ovo–i.e., from nothing. Or at least from 'things' not in the formalized arena of mathematics, such as Hilbert's *"system of things"*. Euclid's definitions of *point* and *line* seem to be whittling these concepts into their pure form from some more materially graspable context (e.g., where lines have breadth)[6] while for Hilbert the essence of *point and line* is their relationship one to the other.

Once one allows the bedrock of—say—Set Theory, definitions are often 'delineations of structure,' cut out by means of quantifiers and predicates but making use of set theoretic, or at least priorly defined objects. E.g. An **abelian group** is a *group* such that ...

But if one uses Set Theory as a 'substrate' on which to build the structures of mathematics, as in the classical *Grundlagen der Mathematik* of Bernays and Hilbert, one must tangle with all the definitional questions that are faced by Set Theory (starting with: *what is a set?* and continuing with the discussion generated by the work of Frege, Russell, etc.). For example, go back to Dedekind's marvelous idea of capturing the notion of *infinite* by discussing self-maps (this notion popularized by people checking into Hilbert's hotel). You might formulate Dedekind's idea this way: a set S is **infinite** if it admits an injective but non-surjective self-map... and then confuse yourself by trying to figure out how this compares with the property that S admits a surjective but non-injective self-map.

And then compare all this with the discussion about the existence of infinite sets in Bernays-Hilbert's *Grundlagen der Mathematik, Vol. I*:

> ... reference to non-mathematical objects can not settle the question whether an infinite manifold exists; the question must be solved within mathematics itself. But how should one make a start with such a solution? At first glance it seems that something impossible is being demanded here: to present infinitely many individuals is impossible in principle; therefore an infinite domain of individuals as such can only be indicated through its structure, i. e., through relations holding among its elements. In other words: a proof must be given that for this domain certain formal relations can be satisfied. The existence of an infinite domain of individuals can not be represented in any other way than through the satisfiability of certain logical formulas...

[6] I want to thank Eva Brann for pointing this out.

The essential roles that 'definition' play for us are: to delineate the objects of interest to be studied; to encapsulate; to abbreviate; and to focus.

As for *focus* consider the difference between *definition* and *characterization*. These are almost synonyms, but—of course—not quite. For example, here are two assertions about an integer $N > 1$.

A. N cannot be expressed as the product of two numbers, each of which is > 1.

B. If N divides the product of two numbers, then N divides one of them.

Now, an integer $N > 1$ satisfies (**B**) if and only if it satisfies (**A**); i.e., if and only if it is a *prime number*:

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

We have three (logically equivalent) choices here. We can proclaim (**A**) to be the *definition* of prime number and (**B**) a logically equivalent *'characterization'* of primeness; we can go the other way taking (**B**) as definition and (**A**) as characterization; or we might simply say that these are two equivalent definitions of prime number.

Our choice determines our focus, but a change in this choice will have no effect in any argument. Given how (**A**) is surely the standard choice, it may be surprising, though, that for structural reasons, in broader contexts, algebraists use (**B**) as their choice for the definition of the concept "prime" rather than (**A**)—and the latter often goes under the name *irreducible*.

Asking Questions:

Often, when one finds the answer to a question that one has been struggling with, one gains the experience to ask a deeper or broader or more fundamental question, or at least to re-ask the original question in a better way. Aristotle tried to classify formats of question-asking (i.e., of answer-seeking) in Book II of the *Physics* or Book V of the *Metaphysics*.

But let's take a more relaxed view and just consider the list:

What?, How?, Where?, When?, By what means?, For what purpose?, etc.

We have the tendency, in Mathematics, to ask *What?* questions. E.g., *what* are all the rational solutions to the equation 4.1 we previously discussed? These can be wonderful questions, and they serve us well, but they mostly have an unstated further mission—to increase our 'understanding' (that dangling gerund!).

In fact, all of the types of questions listed above invite some kind of subtle 'reductionism' to enter into their answers... and even after all of them are answered, there (sometimes) remains a 'final why?'—-i.e., a desire for *more understanding*.

The mathematician Bernard Teissier once pointed out that mathematical language is designed to be adequate for expressing the answers to any of the questions in the above list. But the eureka moments—the moments when you feel you *understand*—when you suddenly "see" something—i.e., the answer to the "why" questions—are not experiences that seem to be faithfully translatable by any utterances at all. Rather, they are inner felt experiences that elude our so carefully constructed language. Nevertheless:

3. Mathematics unified by language

Unifying 'Foundations' and 'Substrate':

Foundational systems for any subject have as aim the establishment of the modus operandi and a language that demarks and unifies the subject. The *Grundlagen der Mathematik* of Bernays and Hilbert certainly had that as purpose and succeeded in its goal. Axiomatic systems with their (related) issues of consistency, and models, are now standard for the setting up of any mathematical discussion: and *Set Theory*—despite its complexities—has long played the role as the *substrate* on which our various structures are built.

These side-questions arise:

- To what extent are such 'Foundations' meant to be explicitly referred to in mathematical practices, and explicit architectural guides for the way we construct our theories?

- To what extent are they meant to be merely in the background, to assure us of the coherence of the structures we develop if and when such assurance is needed?
- To what extent must they even be known by mathematicians as they do their work?

Unifying profiles:

What could be a more basic assertion about the *profile* of Language (meaning: the language we use every day) than to proclaim that there are *nouns* and there are *verbs*; and there are some rules about their interaction.

Category Theory, in the same spirit, offers something of a natural over-view of mathematical theories; or more broadly of any assortment of mathematical objects and their natural interactions.

From the perspective of Category Theory, the *profile* of a mathematical theory consists of its 'objects' $X, Y, \ldots$ and its 'morphisms,' (i.e., the transformations that it allows between its objects) and the (associative) rule that stipulates the composition of morphisms $X \to Y \to Z$ as a morphism $X \to Z$. With the vocabulary of Category Theory, one can then compare different profiles. That is, we can discuss mappings from one category to another (that respect the structures involved) technically: *functors*.

An elementary, but influential, early example displaying the virtue of the category-theoretic vocabulary of *functors* for describing things succinctly is the proof that there is no continuous mapping of the disc D to its boundary S that is the identity mapping when restricted to S. This follows immediately from the classical construction of a functor $\mathcal{H}$ from the category of topological spaces-and-continuous maps to the category of (abelian) groups-and-homomorphisms [7] that sends D to the trivial group and S to a nontrivial group. If there were continuous maps $S \to D \to S$ such that their composition is the identity $S \overset{=}{\to} S$, the functor $\mathcal{H}$ would provide a diagram in the category of groups giving homomorphisms

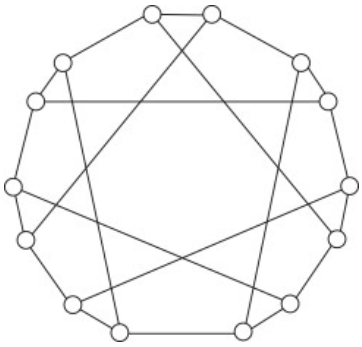
$$\text{nontrivial group} \to \text{trivial group} \to \text{nontrivial group}$$

such that their composition is the identity—which is not possible.

[7] E.g., $\mathcal{H} : X - - \to H_1(X, \mathbb{Z})$

Model Theory begins by offering a format for doing mathematics within an explicitly shaped 'Language' (in the style of 'universal algebra')—where again the substrate is meant to be *Set Theory*—and where its *sentences* interpreted in any 'model' have truth-values that conform to the rules of first-order logic.

The 'opening move' of Model Theory is a powerful and revealing disarticulation of semantics from substance. Here's what I mean: if you are not model-theoretic and want to formulate, say, *graph theory*, you might—for example—just define a **graph** to be given by a set V of vertices and a set E of edges, each edge attaching two distinct vertices and you might also insist that no two vertices are attached by more than one edge. Or you might give a more topological account of this



structure.

In any event, your formulation begins with a set and then some structure is imposed on it.

Model Theory, reverses this. It begins by offering an explicitly shaped language in which first-order logic is incorporated. In the case of our example of graph theory, the language would have a symbol $\mathcal{E}$ labeled as a *binary relation* (symmetric, but not reflexive) in connection with which we label as *true* sentences: $\forall x, y(x\mathcal{E}y \leftrightarrow y\mathcal{E}x)$ and $x\mathcal{E}y \implies x \neq y$. An 'interpretation' of this language—or synonymously, a 'model' for this would be a 'representation' of this language in (some version of) Set Theory. That is, it would give us a set V endowed with a binary relation E for which the labeled-as-true sentences are... in fact true; i.e., such a model is simply a graph, where the set of vertices is the set V and the set of edges is given by the binary relation E.

It is interesting—in thinking about how tightly integrated mathematics is—that there even *is* a clear and consensually agreed upon format of the above sort that can embrace so many aspects of our subject.

4. Mathematics unified by a specific goal

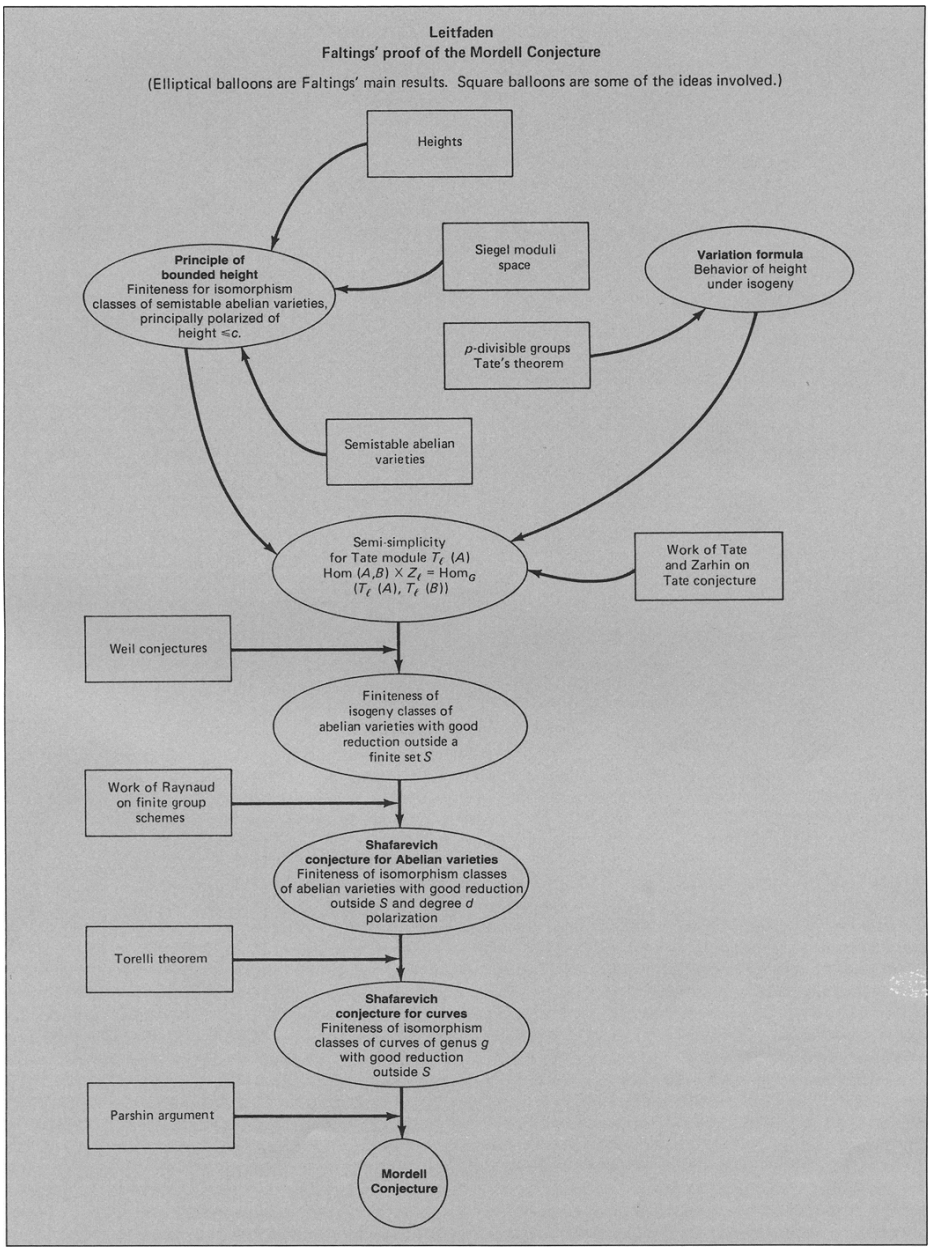
That the equation

$$C : y^2 = x^6 + x^2 + 1 \tag{4.1}$$

we discussed in Section 1 has only one solution in positive rational numbers has been known since 1998. But that it has only finite many rational solutions had been already known at least a decade earlier for it follows from Faltings' Finiteness Theorem—alias: Mordell's Conjecture (that the number of K-rational points on any curve of genus > 1 defined over a number field K is finite). Without going into any discussion of any of the ideas [8], but just to give an overall sense of the number of different concepts that are required for, and are united into, the strategy of *one of* Faltings' proofs of Mordell's Conjecture, take a look at this elegant flow-chart[9] diagramming that proof:

[8] We will return to this later.

[9] created by Spencer Bloch for his article *The Proof of the Mordell Conjecture* that appeared in THE MATHEMATICAL INTELLIGENCER **6** 1984 (41-47)



The striking quality of this flowchart is that it shows how such an intertwined filigree of different mathematical ideas come together to achieve a common mathematical goal. Needless to say, this isn't a unique occurrence, and is, perhaps, more just a normal feature in the history of mathematical progress.

5. Mathematics unified by analogies, and metaphoric connections

. . . man is an analogist and studies relations in all objects.

Emerson; *Nature*, Ch IV on 'Language'

Much has been written about how (or *if*) the agents *metaphor* and *analogy*—being continually at work on the extension and refinement of meaning—are therefore responsible for the broad reach that language enjoys[10].

Somewhat unsettling comments have also been made about analogies—e.g., in their role in mathematics—as in this quotation of André Weil:

> Nothing is more fruitful—all mathematicians know it—than those obscure analogies, those disturbing reflections of one theory in another; those furtive caresses, those inexplicable discords; nothing also gives more pleasure to the researcher. The day comes when the illusion dissolves; the yoked theories reveal their common source before disappearing. As the Gita teaches, one achieves knowledge and indifference at the same time.

Indifference? There are metaphorical bridges, begun in ancient mathematics, that connect subjects and viewpoints cajoling us to view one field from the perspective of another—for example: *geometry as algebra* and *algebra as geometry*. René Descartes, commenting about his merger of algebra and Euclidean geometry, said:

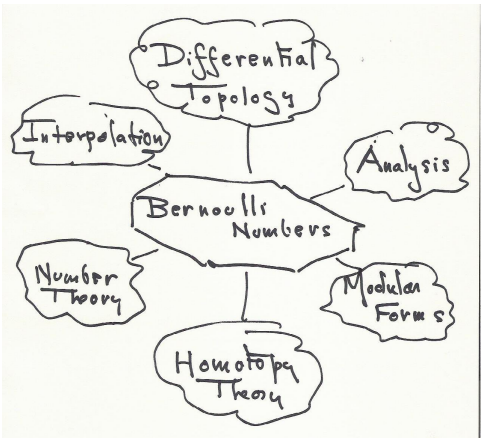
> I would borrow the best of geometry and of algebra and correct all the faults of the one by the other.

This synthesis that mathematicians have created by yoking *geometry* (with its vibrant visual intuition) with *algebra* (with its more verbal, symbolic, combinatorial intuition) is, perhaps one of the most venerable, but hardly the only grand unification of subjects, converting an elusive analogy to an illuminating unity. Mathematics is rife with these. Mathematicians have welded great fields—each with their own distinct

[10] An elaborate discussion of this is in Owen Barfield's *Poetic Diction*, Wesleyan University Press.

brand of intuition—into grand syntheses where there is a combined intuitive power not matched by either alone.

It is very easy to find *connectors* between seemingly disparate areas of research. For example, I once gave a lecture on how such a clean simple notion as *Bernoulli numbers* ties together a constellation of different mathematical subjects—and does it in a way that one actually can experience the profound kinship of these subjects— 'Bernoulli numbers' being the keystone:



Monstrous Moonshine:

An astounding example of a 'connector' between different theories having quite different spirits is what has come to be known as *monstrous moonshine*. It involves a *single* computation, one that any of us can do:

$$196884 = 1 + 196883. \tag{5.1}$$

In the early seventies, the mathematician John McKay made that simple computation, but coupled it with a very important observation. What is peculiar about this formula is that the left-hand-side of the equation, i.e., the number 196884, is well-known to most practitioners of a certain branch of mathematics (*complex analysis*, and the *theory of modular forms*):

- 196884 is the first interesting coefficient of a basic function in that branch of mathematics: the elliptic modular function:

$$j(\tau) = \frac{1}{q} + 744 + \sum_{n\geq 1} c_n q^n = \frac{1}{q} + 744 + \mathbf{196884}q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + \dots$$

while

- 196883 which appears on the right of Equation 5.1 is well-known to most practitioners of (what was in the 1970s) quite a different branch of mathematics (*The theory of finite simple groups*): 196883 is the smallest dimension of a Euclidean space that has the largest sporadic simple group $M =$ (*the Monster group*) as a subgroup of its symmetries: The dimensions n for which M acts irreducibly on an n-dimensional complex vector space–including the 1-dimensional space on which M acts trivially—comprise a finite list of numbers:

(5.2) $$1,\ \mathbf{196883},\ 21296876,\ 842609326, 18538750076, \dots.$$

John McKay went on to make the following puzzling and somewhat amazing observation about a bunch of similar numerical coincidences: the first few Fourier coefficients of the elliptic modular function j can be expressed as sums—with very few summands!—of the dimensions n that appear in the list 5.2. For example:

$196884 = 1 + 196883,$
$21493760 = 1 + 196883 + 21296876,$ and
$864299970 = 2 \times 1 + 2 \times 196883 + 21296876 + 842609326,$
$20245856256 = 3 \times 1 + 3 \times 196883 + 21296876 + 2 \times 842609326 + 18538750076.$

This extremely arresting purely numerical observation suggested a world of new structure: it led to the conjecture that there lurked an infinite sequence of 'natural in some sense' complex representation spaces of the Monster group,

$$V_1, V_2, V_3, \dots, V_n, \dots$$

where, for $n = 1, 2, 3, \dots$ the Fourier coefficient c_n of the elliptic modular function is equal to the dimension of V_n. This seemed, perhaps, so startling at the time that the conjecture

was labeled *monstrous moonshine*. When eventually proved[11] it has given birth to another profound field in mathematics, and intimate links with physics.

The Langlands Program:

In number theory, one of the great analogies that ties together initially very different fields (each with different fundamental guiding intuitions) is called the *Langlands Program*.

The format of the Langlands Program beautifully clarifies the intertwined relationship between the 'local' and the 'global' in arithmetic[12] and representation theory; and proclaims—among other things—a (still largely conjectural) correspondence between two quite different mathematical objects:

Algebraic Number Theory

i.e., the study of (certain) homomorphisms[13] of the Galois group of a number field K **to:**

An algebraic group G

"and this corresponds to"

[11] This involved work of John Conway, Simon P. Norton, Igor Frenkel, James Lepowsky, Arne Meurman and Richard Borcherds.

[12] 'Global' refers to questions about global number fields while 'local' refers to related questions over the finite residue fields coming as quotients of the rings of integers in number fields modulo nonzero prime ideals P, or over P-adic completions. That there is a strong relation between local and global can be seen by considering the classical theorem saying that a positive integer is a square if and only if if is a square mod p for all primes p. That there is some tension between the global and local can be seen by the classical homogenous cubic example due to Selmer: $3x^3 + 4y^3 + 5z^3 = 0$, which has no nontrivial rational integer solutions but does have nontrivial p-adic solutions for all primes p.

[13] more explicitly: certain irreducible *Representations ρ (unramified at all but finitely many places) of the Galois group of a number field K* into the group of complex, or ℓ-adic, points of GL_n (or more generally of some reductive algebraic group G)

Analysis; Automorphic Representations

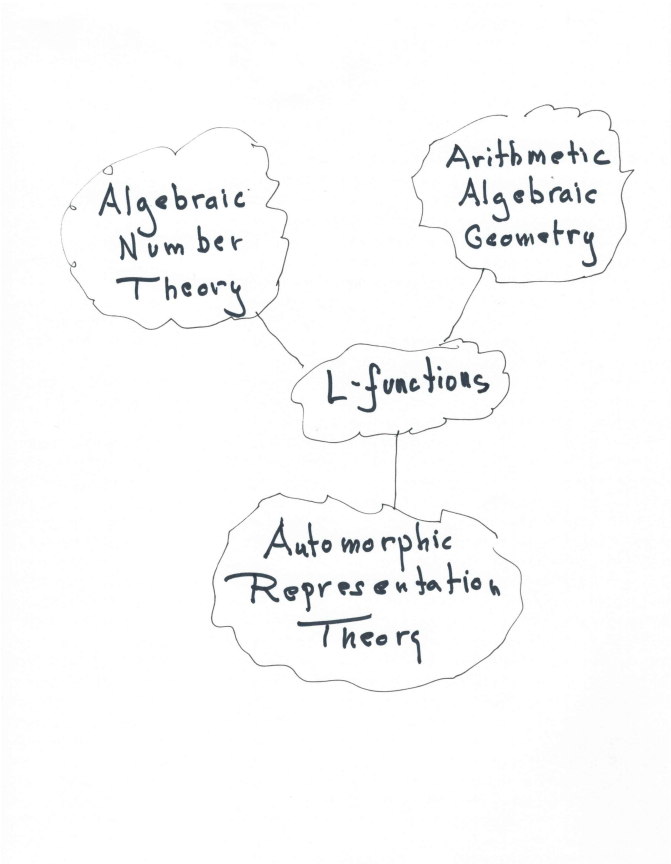
i.e., the study of (certain usually infinite dimensional) ('automorphic') representations **of:**

An algebraic group G

In the correspondence alluded to above, the 'blue' group and the 'red' group are specifically related: the 'blue' group is *the Langlands dual* of the 'red' group.

What ties things together is the theory of L-functions. The L-functions that enter into this story are Dirichlet series belonging to what is called **the Selberg class of L-functions**: these satisfy axioms that guarantee that they—as Dirichlet series—are absolutely convergent in some right half-plane; they have an infinite product expansion indexed over prime numbers; they extend by analytic continuation to a function on the entire plane—analytic, or meromorphic with only a simple pole at $s = 1$; and they satisfy a functional equation of a prescribed type.

A diagram in the style of the "Bernoulli number diagram" in Section 5 above that gives a visual sense of the position of those fields as conjoined by the Langlands program would be:



Note: *Arithmetic Algebraic Geometry* enters in this story significantly insofar as the major way of getting compatible families ρ of Galois representations is by considering the Galois action on the étale cohomology of algebraic varieties defined over number fields.

The classical theory of modular forms enters the story in that they give rise to 'automorphic representations' π.

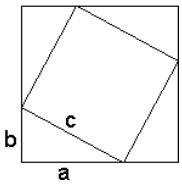
To see the *tiniest* hint of the flavor of this correspondence, see the Appendix, Section 10 below.

6. Many proofs of the 'same' result

The Pythagorean Theorem can be thought of as a *theme* that unifies the multitude of its different proofs. Each of these proofs *verify* the same statement ($a^2 + b^2 = c^2$) about a right-angle triangle Δ but they

may *explain* or perhaps just *illustrate* ever-so-slightly different aspects of it. E.g., Euclid's Book I Prop. 47 cuts the square built on the hypotenuse of Δ to construct two rectangles each equal in area to the square built on the two sides; while Euclid's Proposition VI.31 decomposes Δ in the appropriate proportions. Or, there is the proof starting with a square with sides of length $a + b$, and removing four pieces isomorphic to Δ to get the square built on the hypotenuse:

$$(a + b)^2 - 2ab = c^2.$$



This proof seems to go back to the 12th century Hindu mathematician Bhaskara. (It may also have an—independent?— early Chinese origin.) The special case when $a = b$ (showing that a square—call it B—with sides of length equal to the the diagonal of another square—call it A– is twice the area of A) is the mathematical exercise in Plato's *Meno*.

I wonder whether anyone has dealt with the full gamut—i.e., the hundreds—of proofs of the Pythagorean Theorem to see if–taken all together—they form some sort of integral, and graspable, structure.

. . . and is there a similar integral, graspable structure for the different ways of showing finiteness of K-rational points?

7. Different strategies to prove finiteness of K-rational points

It may be instructive to think similarly about the different ways of achieving the *finiteness statements* that the different proofs of Mordell's Conjecture offer us [14].

[14] For a marvelous exposition of various approaches to Diophantine finiteness results (focusing more on issues related to Roth's Theorem, Thue-Siegel Roth, and Vojta's approach—and a bit on Faltings' approach) see Michael Nakamaye, *Roth's theorem: an introduction to diophantine approximation. Rational points, rational curves, and entire holomorphic curves on projective varieties* Contemp. Math., 654, Centre Rech. Math. Proc., Amer. Math. Soc., Providence, RI, (2015) 75-108.

Given that this is a lecture about the unity of mathematics, and given that there are (at present) three strikingly different approaches to proving finiteness of K-rational points, it might seem as if I had the obligation to at least try—in some way or other—to unify these approaches. Success in this has—so far—eluded me. The three different approaches prove three very different finiteness results each of which has, as corollary finiteness of K-rational points.

The curious list:

(i) **Gerd Faltings' first proof:-**establishing finiteness of the number of isomorphism classes of abelian varieties over a number field of fixed dimension and of good reduction outside a fixed finite set of primes of that number field.

(ii) **Paul Vojta's proof:-**(with Elaborations and some simplifications by Faltings, Bombieri, McQuillan) where if given a vast number m of rational points on a given curve one manages to find a line bundle on the m-th power of C and a section that doesn't vanish on the m-tuple formed by those m rational points but does vanish on *such a nearby divisor* that... leads to a contradiction. This directly follows the format of the classical proof of Roth's Theorem.

(iii) **The p-adic methods of Chabauty:**–available only under some hypotheses—made more explicit by Robert Coleman, and most recently extended significantly by Kim and others. Here is what is obtained by this method, when the hypotheses are met: For some primes v of the number field one constructs a v-adic (locally) analytic function ϕ on the curve C over the completion, K_v, of K at v which has two properties:

  (a) The (locally) analytic function ϕ has only finitely many zeroes on $C(K_v)$.

  (b) ϕ vanishes on every K-valued point of C.

  In effect, this approach, when C and K meet its hypotheses, offers what one might call an effective 'v-adically enhanced' upper bound on the number of rational points, since it comes with some kind of v-adic extra information: the v-adic function ϕ is often obtainable effectively—i.e., its power series at residue

discs are computable modulo p^N for any N—so one can effectively determine an upper bound for the number of zeroes—and a finite set of discs of however small radius one wishes which contain whatever K-rational points the curve has. It is an interesting question, I think, to capture the exact nature of 'effectiveness' in this method.

The appendix contains some more commentary about these three—significantly different—methods of achieving finiteness.

8. Unity

So what is it that unifies a subject?

An articulable goal? This might not apply to mathematics, a subject that is constantly expanding its terrain and ambition.

A common language? In the story of Diophantus's Problem, as we've seen, a precise response is given to a precise issue raised ∼1700 years ago, and the *statements* of both issue and response—if not the vocabulary of the methods used—are expressed essentially in the same language.

A common origin? But this neither addresses, nor predicts, the powerful forces that keep the subject tightly coherent as it progresses.

A few of these forces we've reviewed in this essay, but how could we even begin to give an adequate account of the vast number of surprise inner connections our subject has[15]?

I think, though, that the main overarching 'unifier'—and I know I might have to defend this!—is that we all happily share a general sense of what constitutes *mathematical demonstration*. By this I don't mean that we have fixed a precise language, and foundations for our subject that regulates what stands for mathematical proof. I mean that, even before doing that, we already have a general sense and conviction of what *should count* as mathematical demonstration, and it is this inner sense that is the bedrock of our subject, and that induces us to go about trying to fix such a language and foundation.

Some years ago, Noah Feldman and I ran a seminar course—*The Nature of Evidence*—in the Harvard Law School, where we asked experts in different fields (Art History, Economics, Physics, Biology, Mathematics, Law) to offer us reading lists and presentations to allow us to

[15] Some of these connections are not only surprises but quite difficult to classify such as the wonderful way in which Charles Fefferman used the solution to the Kakeya Problem to establish an important feature of L^2-norm that distinguishes is from L^p-norm for $p \neq 2$.

learn what constituted *evidence* in their field. What does one have to provide to guarantee general agreement among the experts in—say—Economics that some new result has been established in their field? It is natural—but still it was striking to me— that each field demands a markedly distinct structure for what it accepts as evidence. And if there are internal disagreements, (or let us say, 'discussions') these discussions reinforce the sense that the precise nature of admissible evidence in the field uniquely pinpoints the field.

We have been contemplating a few of the many features that help unify Mathematics—help establish it as a coherent project. But it seems to me that—despite my falling into a clear vortex of circular reasoning—it is the unique nature of—a universally available—mathematical sensibility that 'holds it together.'

9. Appendix: Further commentary on the three methods of obtaining finiteness of the number of K-rational points in a curve of genus > 1 over a number field K

A. Faltings' First Proof. The rough frame for the string of finiteness assertions in the diagram in Section 4 (that charts Faltings' initial proof that a curve C of genus > 1 over a number field K has only finitely many K-rational points) is quite complex, but consists of three basic elements (working up from the bottom in Spencer Bloch's diagram we discussed).

(i) The construction by Parshin of a certain *sufficiently varying* family[16] of curves C'_x (of a certain fixed genus depending only on C) parametrized by the points x of the original curve C. (The family is defined over a finite extension K'/K, and there is a finite set S' of primes of K' outside of which C'_x has good reduction where x ranges over all K'-valued points of C.) So we have the basic association:

$$x \mapsto C_x$$

[16] See also the recent result regarding finiteness of K-rational points on a curve C—*Diophantine Problems and the p-adic Torelli Map* due to Lawrence and Venkatesh—where the starting construction is of a family of algebraic varieties over C, and the major issue is that it be *sufficiently varying*.

of K' rational points x of C to curves C_x (over K') of a fixed genus of good reduction outside a fixed finite set of primes of K'. Moreover, this mapping is *finite-to-one.*

(ii) The fact that the natural association

$$\mathcal{C} \mapsto Jac(\mathcal{C}) \mapsto \tilde{J}ac(\mathcal{C})$$

that sends the set of isomorphism classes of curves $\mathcal{C}$ of fixed genus defined over a number field K' and having good reduction outside a finite set S' of primes of K' to $Jac(\mathcal{C})$, the isomorphism class of their jacobian, and thence to $\tilde{J}ac(\mathcal{C})$, the corresponding isogeny class, is *finite-to-one.*

(iii) The clincher then being that for a given number field K' and finite set S' of primes of K' and for a fixed dimension g' there are only finitely many isomorphism classes (or isogeny classes)over K' of abelian varieties of dimension g' with good reduction outside S'.

B. Paul Vojta's method. Guided by analogies that the arithmetic features that the problem of bounding numbers of rational points have with two other arenas of problems—*hyperbolic geometry* and *approximation of algebraic numbers by rational numbers*—Vojta established finiteness by a very beautiful route. (Elaborations and some simplifications were also subsequently given by Faltings, Bombieri, McQuillan.) One way to get a feel for Vojta's strategy is to think of the shape of the proof of the classical Roth's Theorem. Roth's Theorem guarantees that for a given algebraic irrational (real) number α, and for any $\epsilon > 0$, there are only finite many rational numbers p/q such that

$$|\alpha - p/q| < \frac{1}{|q|^{2+\epsilon}}.$$

The strategy is to pass to high dimension: if, say, for a given such α and ϵ, there were infinitely many such rational numbers p/q, then for any number m there would be m such rational numbers p_i/q_i and we could choose them so that the denominators q_i are growing very rapidly (as rapidly as necessary for this ensuing strategy to work!). So choose such a collection of m rational numbers. We want to view the m-tuple $(p_1/q_1, p_2/q_2, \ldots, p_m q_m)$ as a pretty good approximation to the diagonal m-tuple $(\alpha, \alpha, \ldots, \alpha) \in R^m$.

Now one starts looking for what is called an *auxiliary polynomial* (although, as you'll see, *auxiliary* may be a bit too demeaning a term). Look among polynomials $f(x_1, x_2, \ldots, x_m)$ with coefficients in $\mathbb{Z}$ (and multi-degree $(d_1, d_2, \ldots, d_m)$; here d_i is the degree of x_i in $f(x_1, x_2, \ldots, x_m)$) where the multidegrees are required to have some relation to the denominators q_i; namely $d_i \cdot \log q_i$ should be roughly constant. Moreover, $f(x_1, x_2, \ldots, x_m)$ is also required to have two properties:

(i) The polynomial $f(x_1, x_2, \ldots, x_m)$ doesn't vanish at $(p_1/q_1, p_2/q_2, \ldots, p_m/q_m)$.

(ii) The polynomial $f(x_1, x_2, \ldots, x_m)$ does 'vanish sufficiently' at the point $(\alpha, \alpha, \ldots, \alpha)$.

I haven't said what "roughly constant" means in the above paragraph, nor 'vanish sufficiently' in (b). But the key to the method is to get contradictory upper and lower bounds for $|f(p_1/q_1, p_2/q_2, \ldots, p_m/q_m)|$ if m is sufficiently large. Condition (a) immediately implies that

$$|f(p_1/q_1, p_2/q_2, \ldots, p_m/q_m)| \geq \frac{1}{\prod_i q_i^{d_i}},$$

which gets us a lower bound, and condition (b)—together with the fact that $(p_1/q_1, p_2/q_2, \ldots, p_m/q_m)$ is a pretty good approximation to $(\alpha, \alpha, \ldots, \alpha)$ gets us an upper bound—and these bounds, taken together form a contradiction if m is sufficiently large[17]

The beauty of Vojta's method is that it transforms this strategy to an analogous strategy establishing finiteness of rational points over number fields for curves of genus ≥ 2. The judicious choice of the sequence of m approximants p_i/q_i in the proof of Roth's Theorem is replaced—by Vojta—by a judicious choice of of m rational points on the curve C that is being studied, leading us to focus on certain line bundles on the m-th power of C. The choice of auxiliary function in Roth's Theorem is replace by a judicious choice of a section of those line bundles, leading to two contradictory bounds, as above.

[17] One even has a constructive upper bound due to Davenport and Roth: See their *Rational approximations to algebraic numbers* online: `https://doi.org/10.1112/S0025579300000814`. See also the related:

- Joseph Silverman's *A quantitative version of Siegel's theorem: Integral points on elliptic curves and Catalan curves* J. Reine Angew. Math. **378** (1987), 60-100, and
- J.-H. Evertse and H. P. Schlickewei, *A quantitative version of the absolute subspace theorem* J. Reine Angew. Math. **548** (2002), 21-127.

C. Claude Chabauty's method. The method of Chabauty proves finiteness under a somewhat restrictive assumption, namely that the rank of the Mordell-Weil group over K of J is *strictly less than* the dimension of the J—i.e., of the genus of C. Choosing an appropriate prime v of K lying above a rational prime p, note that because of elementary consideration of rank—given our hypothesis—$J(K)$ must lie in a p-adic Lie group $\mathcal{L}$ of codimension ≥ 1 in the p-adic Lie group $J(K_v)$ — and yet the p-adic curve $C(K_v)$ *cannot* be entirely contained in such a p-adic Lie group $\mathcal{L}$. One shows, then, that the intersection of p-adic analytic manifolds $C(K_v) \cap \mathcal{L}$ is *finite*. Since this intersection contains $C(K)$ one concludes the finiteness one has sought.

C.1. *Robert Coleman's version of Chabauty's method.* Robert Coleman revisited Chabauty's method by capturing the intersection $C(K_v) \cap \mathcal{L}$ as the zero-locus of a constructed p-adic differential on C, thereby giving, apparently, relatively good bounds of finiteness allowing one—in certain instances—to actually determine the set of rational points of the curve (granted: Coleman's method also works only in case the rank of Mordell-Weil of the jacobian is strictly less than the genus of the curve).

C.2. *Minhyong Kim's extension of Coleman-Chabauty.* More recently, though, Minhyong Kim revisited Chabauty-Coleman's method by replacing the mapping of a curve to its jacobian by an elegant refinement mapping to spaces of torsors for appropriate unipotent groups. The result is a powerful extension of the effective nature of Chabauty-Coleman's method that led to—among many other things— an effective determination[18] of the number of $\mathbb{Q}$-rational points on any curve C that has the property that its jacobian J is isogenous to a product of two (positive dimensional) abelian varieties over $\mathbb{Q}$ and the Mordell-Weil rank of J is *equal to* the genus of C. These conditions are indeed fulfilled by the equation 4.1 with which this essay began[19].

[18]Theorem 1.1 of *An effective Chabauty-Kim theorem* by Jennifer Balakrishnan, Netan Dogra arxiv.org/abs/1803.10102v2

[19]The curve C, being of genus 2, has only finitely many rational points, by Faltings' classical theorem, but as for the actual determination of that finite set of points, the classical Chabauty-Coleman method for determination of rational solutions does not apply since the Mordell-Weil rank of its jacobian is *equal to* (not *strictly less than*) the genus of C: the quotient of C by the involution $(x, y) \mapsto (-x, y)$ is the elliptic curve 496a1 in Cremona's table; the quotient by $(x, y) \mapsto (-x, -y)$ is 248a1. Both these elliptic curves have Mordell-Weil rank 1 so the

C.3. *Summary.* To return to the question of different ways of achieving the *finiteness statements* that the different proofs of Mordell's Conjecture offer us:

The last three listed methods achieve their finiteness conclusion in the following general way. If C denotes the curve over K for which the method applies, then for some prime v of K one constructs a certain (locally) v-analytic mapping $\omega : C(K_v) \to V(K_v)$ where V is a certain v-analytic variety (e.g., ω might be the Abel-Jacobi mapping to the jacobian of C, or it might be a section of some line bundle) such that there is a specified v-analytic subvariety $V_o \subset V$ such that the pullback $\omega^{-1}(V_o(K_v)) \subset C(K_v)$ is **(a)** finite and **(b)** contains $C(K)$. This generally gives a constructed upper bound for the number of K-rational points, as well as v-adic information about them[20].

And although there is hint of a resonance between the method of Faltings' first proof and the method of these last three, it would be interesting—it seems to me—if one could give a single convincing description of a 'way of achieving finiteness' that encompasses them. Even more challenging would be to describe the method of Vojta (or its elaborations by Faltings, Bombieri, MicQuillen, and Nakamaye) in terms that illuminates its relationship to the other methods.

10. Appendix: A hint of the flavor of the Langlands Correspondence

One sees a speck of this grand correspondence already in the classical set-up regarding the Galois group (over $K = \mathbb{Q}$) of the cyclotomic field $L_n := \mathbb{Q}[e^{2\pi i/n}]$. This is the number field obtained by adjoining the 'primitive' n-th root of unity $\zeta_n := e^{2\pi i/n}$ to the field $\mathbb{Q}$. Putting

$$\Gamma_n := \mathrm{Gal}(L_n/\mathbb{Q})$$

Jacobian of C, which is isogenous to the product of those two elliptic curves, has Mordell-Weil rank 2. Nevertheless, and happily, recent extensions of Chabauty-Coleman's method are available—e.g., the Theorem 1.1 referred to in the previous footnote—and such methods are used in Wetherell's thesis to establish that $x = \frac{1}{2}$ is the unique positive rational solution to 4.1.

[20] The very recent preprint of Brian Lawrence and Akshay Venkatesh *Diophantine Problems and the p-adic Torelli Map* proves (in certain cases) finiteness of rational points on curves by a somewhat different method which nevertheless fits the above general description. The non-archimedean places v play the major role in all these cases. In contrast, is there some context where one learns finiteness of K-rational points by finding those K-rational points as (for example) a subset of the zeroes of some constructed complex analytic function, or section of some line bundle over C (over $\mathbb{C}$)?

one has the fundamental *canonical isomorphism* (due, essentially—and in different language) to Gauss:

$$GL_1(\mathbb{Z}/n\mathbb{Z}) \quad = \quad (\mathbb{Z}/n\mathbb{Z})^* \quad \xrightarrow{\sigma} \quad \Gamma_n. \tag{10.1}$$

Here $(\mathbb{Z}/n\mathbb{Z})^*$ is the group of units in the ring of integers $\mathbb{Z}/n\mathbb{Z}$ in the Galois group, which we could also think of as the group of $\mathbb{Z}/n\mathbb{Z}$-valued points of the algebraic group GL_1. The isomorphism σ is given by the rule: $a \mapsto \sigma(a) : L_n \to L_n$ where for an integer a prime to n, $\sigma(a)$ is the (unique) field-automorphism of L_n that sends $e^{2\pi i/n}$ to $e^{2\pi ia/n}$—and in fact raises every n-th root of unity in L_n to the a-th power.

The groups in Equation 10.1 being abelian, their irreducible (say complex) representations are just multiplicative $\mathbb{C}^*$-valued characters, and the isomorphisms in 10.1 then can be viewed as 'correspondence' that identifies characters $\rho : \Gamma_n \to \mathbb{C}^*$ with classical Dirichlet characters

$$\chi : (\mathbb{Z}/n\mathbb{Z})^* \to \mathbb{C}^*$$

or equivalently, (even though it may be too pedantic to view it this way) with characters on the $\mathbb{Z}/n\mathbb{Z}$-valued points of the algebraic group GL_1.

The connection between χ and ρ in Equation 10.1 is mirrored by the equality of their relevant L-functions, the **Dirichlet L-function** $L(\chi, s)$ and the **Artin L-function** $L(\rho, s)$:

$$L(\chi, s) = L(\rho, s), \tag{10.2}$$

where the **Dirichlet L-function** is:

$$L(\chi, s) := \sum_k \chi(k)k^{-s} = \prod_{p \not| n} (1 - \frac{\chi(p)}{p^s})^{-1} \tag{10.3}$$

(where $\chi(k) = 0$ if $\gcd(k, n) > 1$)

and the **Artin L-function** is:

$$L(\rho, s) := \prod_{p \not| n} (1 - \frac{\rho(Frob_p)}{p^s})^{-1} \tag{10.4}$$

(where $Frob_p \in \Gamma_n$ is the "Frobenius element at p.")

The formulation of the L-functions in Equations 10.3 and 10.4 as infinite products of 'local factors' associated to (all but finitely many) primes p points to the task that L-functions have: to tie local and global properties together.

The equality 10.2 follows since $Frob_p(\zeta_n) = \zeta_n^p$ so the correspondence 10.1 identifies $Frob_p$ with the image of p in $(\mathbb{Z}/n\mathbb{Z})^*$ for primes $p \not| n$.

Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

E-mail address: mazur@math.harvard.edu