THE BULLETIN OF SYMBOLIC LOGIC
Volume 16, Number 3, Sept. 2010

WHAT DOES IT TAKE TO PROVE FERMAT'S LAST THEOREM? GROTHENDIECK AND THE LOGIC OF NUMBER THEORY

COLIN MCLARTY

Abstract. This paper explores the set theoretic assumptions used in the current published proof of Fermat's Last Theorem, how these assumptions figure in the methods Wiles uses, and the currently known prospects for a proof using weaker assumptions.

Does the proof of Fermat's Last Theorem (FLT) go beyond Zermelo Fraenkel set theory (ZFC)? Or does it merely use Peano Arithmetic (PA) or some weaker fragment of that? The answers depend on what is meant by "proof" and "use," and are not entirely known. This paper surveys the current state of these questions and briefly sketches the methods of *cohomological number theory* used in the existing proof.

The existing proof of FLT is Wiles [1995] plus improvements that do not yet change its character. Far from self-contained it has vast prerequisites merely introduced in the 500 pages of [Cornell et al., 1997]. We will say that the assumptions explicitly used in proofs that Wiles cites as steps in his own are "used in fact in the published proof." It is currently unknown what assumptions are "used in principle" in the sense of being proof-theoretically indispensable to FLT. Certainly much less than ZFC is used in principle, probably nothing beyond PA, and perhaps much less than that.

The oddly contentious issue is *universes*, often called *Grothendieck universes*.[1] On ZFC foundations a universe is an uncountable transitive set U such that $\langle U, \in\rangle$ satisfies the ZFC axioms in the nicest way: it contains the powerset of each of its elements, and for any function from an element of U to U the range is also an element of U. This is much stronger than merely saying $\langle U, \in\rangle$ satisfies the ZFC axioms. We do not merely say the powerset axiom "every set has a powerset" is true with all quantifiers relativized to U. Rather, we require "for every set $x \in U$, the powerset of x is also in U" where

Received August 16, 2009.

I thank Jeremy Avigad, Angus Macintyre, Barry Mazur, and an anonymous referee for advice, which is certainly not to say any of them agrees with everything here.

[1]See Grothendieck [1971] and the fuller account Artin et al., [1972, vol. I, pp. 185–217]. We abbreviate these books as SGA 1 and SGA 4 respectively.

1079-8986/10/1603-0003/$2.90

no quantifier in the definition of the powerset of x is relativized to U. What looks like the powerset of x as seen from inside U has to *be* the powerset as seen in the larger ambient world of sets. The condition on images of functions is similarly stronger than saying $\langle U, \in \rangle$ satisfies the replacement axiom scheme relativized to U. It says every function from an element of U to U which exists in the ambient world of sets is itself an element of U. This extra strength guarantees that any set theoretic construction applied to sets in U will give the same result whether it is interpreted inside of U or in the larger ambient world of sets. The use of universes constantly depends on that.

Grothendieck gave a proof of what set theorists already knew: the definition of a universe in ZFC is the same as saying U is the set V_α of all sets with rank below α for some uncountable strongly inaccessible cardinal α [Artin et al., 1972, vol. I, p. 196]. Since each universe models ZFC, the existence of a universe or of an uncountable strongly inaccessible cardinal is not provable in ZFC. Grothendieck's own axiom of universes posited that every set is contained in some universe, which by replacement implies there are proper-class many successively larger universes corresponding to successively larger inaccessibles. We write ZFC+U for the more modest theory of one universe. That is, ZFC+U consists of ZFC plus the assumption of a universe, or equivalently the assumption of one uncountable strongly inaccessible cardinal.

So ZFC+U certainly implies more statements of arithmetic than ZFC alone.[2] This is Gödel's observation: any axiom implying consistency of ZFC thereby implies statements of arithmetic that ZFC does not, since consistency of ZFC can be expressed as a statement of arithmetic which is not implied by ZFC. This makes the assumption of a universe quite different from the continuum hypothesis or other axioms which extend ZFC without increasing the consistency strength or otherwise implying any new arithmetic. But we will see this Gödel phenomenon does not bear on FLT.

This paper aims to explain how and why three facts coexist:

1. Universes organize a context for the rather explicit arithmetic calculations proving FLT or other number theory.
2. Universes can be eliminated in favor of ZFC by known devices though this is never actually done (and this remains far stronger than PA).
3. The great proofs in cohomological number theory, such as Wiles [1995] or Deligne [1974], or Faltings [1983], use universes in fact.

Large cardinals as such were neither interesting nor problematic to Grothendieck and this paper shares his view. For him they were merely legitimate means to something else. He wanted to organize explicit calculational arithmetic into a geometric conceptual order. He found ways to do this in *cohomology* and used them to produce calculations which had eluded a decade

[2] Throughout this paper "arithmetic" means first order arithmetic, that is statements of PA either in the language of PA or interpreted in ZFC.

of top mathematicians pursuing the Weil conjectures [Osserman, 2008]. He thereby produced the basis of most current algebraic geometry and not only the parts bearing on arithmetic. His cohomology rests on universes but weaker foundations also suffice at the loss of some of the desired conceptual order.

Section 1 gives one specific, central use of universes in Wiles's proof of FLT. Section 2 introduces prospects for proving FLT in PA or a weaker arithmetic. Sections 3–4 sketch cohomological number theory and Grothendieck's strategy. Large structures occupy Sections 5–7, including comparison of three successively stronger extensions of ZFC at the end of Section 5. We cite Deligne [1977, 1998] to show there is no contradiction in finding universes dispensable in principle and useful in fact. Both are true. There is a want of perception in denying either one. At the end of Section 7 we describe what is currently known about expressing cohomological proofs in ZFC without universes. It can certainly be done with some loss to the theoretical organization and we give a conjectural way to do it with no perceptible loss. Section 8 revisits the question of bringing Wiles's proof closer to PA. No one who has looked at Wiles's proof seriously doubts that it could be unwound into a rather high order non-conservative extension of PA, say 8-th order, by perfectly routine means which however would do tremendous damage to the theoretical organization. There is some evidence that a great deal of progress in arithmetic, not routine, can produce a version of the proof in a conservative higher order extension of PA and thus effectively in PA. We can set no limit to how far the proof may be simplified by currently unforeseeable progress in arithmetic. Section 9 draws conclusions for the foundations of mathematics.

There are two quite separate size ranges in this paper. We call a set or structure "large" if it is at least as large as some universe. What we call "very small" structures are all at most continuum sized. Nearly every specific structure we talk about is either large or very small in this sense.

§1. Use of universes. Harvey Friedman offers a clear and simple statement: "I have been told that there is absolutely no trace back from the references used in the body of the Wiles paper to universes."[3] But we will see in a moment that a key reference in Wiles's proof goes straight back to Grothendieck and universes.

The same or another unnamed expert dismisses any practical role for universes: "Nobody who understands such proofs does anything but think about very small structures from the start till the end."[4] This is true most

[3]E-mail list FOM, Friedman post titled "`Report from expert, Tue Apr 6, 1999`" archived at `cs.nyu.edu/pipermail/fom`.

[4]E-mail list FOM, Friedman post titled "`Using Universes? The expert speaks again, Thu Apr 8, 1999`" archived at `cs.nyu.edu/pipermail/fom`.

of the time and is central to cohomological number theory. Grothendieck created the large structures of cohomology to work so smoothly that one reaches arithmetic through them almost without thinking about them.

When beginning active research a number theorist may be well advised to become familiar with theorems on large structures from SGA 4 and related references before lingering long to study the proofs—and rather focus on arithmetic and geometry—until the theorist needs to prove some modified form of one. Barry Mazur pointed out this strategy to me while stressing that anyone who actually works with the ideas will modify many general results over time so that mastery requires both the large-structure theorems and a great deal of small-structure arithmetic. In fact people who understand the proofs routinely cite the large-structure theorems in print, and not rarely re-work the proofs to cover new cases.

Wiles explains how his search for the proof was blocked at one point by a specific problem of arithmetic. He says "the turning point in this and indeed in the whole proof came" when the search led him to two cohomology invariants and "I learned that it followed from Tate's account of Grothendieck duality theory for complete intersections that these two invariants were equal" [Wiles, 1995, p. 451]. The body of the proof (pp. 486–7) cites a source:

> For a summary of the duality statements used in this context, see [Mazur, 1977, §II.3] . . . . To justify the reduction in detail see the arguments in [Mazur, 1977, §II.3].

Mazur does not give complete proofs but cites Grothendieck and Dieudonné [1961] which we abbreviate EGA III, and Deligne and Rapoport [1973] which cites the same parts of EGA III. Grothendieck and Dieudonné use functor categories between locally small categories (p. 349). From the viewpoint of ZFC these locally small categories are proper classes. A function between proper classes is a proper class, so any "set" of functions between two proper classes is a collection of proper classes. Let us call such a collection a *superclass*.

If we are concerned to rise in rank as little as possible then by apt choice of details we can say in ZFC+U locally small categories have the rank of the universe U and the functor categories between them are superclasses one rank higher.[5] Limiting the rise in rank this way is actually pointless in ZFC+U which has sets of rank β above U for every ordinal number β; but it would be crucial if we wanted to use weaker extensions of ZFC which only add proper classes and a limited number of ranks above them. Anyway that will not be the end of our rise in rank, since categorical manipulation of these superclass categories means placing them into categories of yet

[5] Define a category as a composition operator on a class of arrows. If you defined a category as a Kuratowski ordered pair of a class of arrows and a composition operator on it then locally small categories would be two ranks above U.

higher rank. Grothendieck's foundation for this was universes as he says on the first page of chapter 0 of the book version of EGA [Grothendieck and Dieudonné, 1971, p. 19].

Wiles [1995] and Mazur [1977] focus their attention on very small structures. That is finite or countable or at most continuum-sized structures. But they apply general theorems which Grothendieck and Dieudonné [1961] prove by placing those very small structures inside large structures. These theorems are still widely cited today and are still proved using universes as by Lipman and Hashimoto [2009, pp. 160, 287].

Deligne and Rapoport explain that "these techniques give systematic means" of presenting and proving their results (p. 151). All the authors including Grothendieck and Dieudonné know that fragments of these theorems adequate to any given application in arithmetic can be stated inside ZFC by abandoning the system in favor of technicalities and circumlocutions suited to that application. The authors prefer systematic means since the material is lengthy enough already.

§2. Prospects for a weak proof. Angus Macintyre [forthcoming] lays out a program to express the Modularity Thesis (MT) central to Wiles [1995] as a Π^0_1 statement of arithmetic and argues that it is provable in PA. This program could lead to a PA proof of MT, and possibly one of FLT without using MT. It calls for substantial new work in arithmetic. While closely based on Wiles [1995] it is no routine adaptation.

Macintyre points out that analytic or topological structures such as the p-adic, real, and complex numbers enter Wiles's proof precisely as completions of structures such as the ring of integers, or the field of rational numbers, which are interpretable in PA. Macintyre outlines how to replace many uses of completions in the proof by finite approximations within PA. He shows how substantial known results in arithmetic and model theory yield approximations suited to some cases. He specifies other cases which will need numerical bounds which are not yet known. Theorems of this kind can be very hard. He notes that even routine cases can be so extensive that "it would be useful to have some metatheorems" (page 14). The program will be a huge amount of work. If it succeeds by using metatheorems then further progress in the same direction might or might not eventually allow us to eliminate the metatheorems in favor of an explicit proof in PA.

The situation is familiar: A more elementary proof normally requires more delicate calculations. These normally reveal more information. It is normally longer. We may learn a great deal by showing it exists even in cases where it would be idle labor to make it explicit. The special difficulty for FLT is the size of the currently known proof. Wiles [1995] gives 84 references. Many prove steps that Wiles needs. Most are not short themselves. And they rely on other quite advanced references.

Harvey Friedman conjectures that FLT is provable in Exponential Function Arithmetic (EFA). See Avigad [2003].[6] There is currently no independent strategy to prove this. Probably the most promising is to begin with Macintyre's program and, so far as that succeeds, try to take it on to EFA.

The point for us here applies to EFA much more strongly than to PA just because EFA is much weaker proof-theoretically. We can put it in the terms of Avigad [2003, p. 270]. Someone might some day give "an informal proof that there is a formal derivation of the theorem in some conservative extension of EFA" without being able to give any independent "informal description of a derivation of the theorem in EFA." Our terminology would describe that situation by saying the conservative extension is used in fact and EFA is used in principle. We would know EFA proofs of FLT exist while that knowledge would depend on actually describing a proof in the conservative extension.

Some have identified me as a "proponent" of Grothendieck's methods, and conclude that I oppose finding proofs from weak principles. The premise is roughly true, the conclusion is absurd. I am not against finding any proofs. But it will always remain that the first known proof of FLT and the only one yet known today fifteen years after Wiles gave it, cites and relies on published proofs that explicitly use universes. Nor does Macintyre in any way object to this proof! There is synergy and not opposition between Macintyre's program and Grothendieck's methods. Macintyre [2003] repeatedly urges model theorists to look more at those methods.[7]

It would be fantastic to find the weakest comprehension and induction principles sufficient for FLT but graduate number theory seminars will probably not teach those principles. If the history of mathematics is any guide, then we can be sure that over time the proof of FLT will be radically simplified but that is not the same thing as looking for the weakest logical principles. For the foreseeable future it is likely that any proofs of FLT to be found in weak theories of arithmetic will be discovered in the first place, and will be comprehensible after they are discovered, only by applying metatheorems to some shorter known proof using stronger logic. In this context Wiles [1995] counts as a short proof.

§3. The idea of cohomological number theory. The past 50 years have seen huge progress in number theory in the form of *arithmetic algebraic geometry* using certain spaces called *schemes*. The project goes back to Riemann, Dedekind, and Kronecker assimilating algebraic numbers to algebraic functions on a Riemann surface but it relies on tools of cohomology created by Serre and Grothendieck in the 1950s. Schemes are no harder to define

[6]EFA is first order arithmetic allowing only quantifier free induction, taking successor, addition, multiplication, and exponentiation as binary operators.

[7]E.g., topos theory on p. 197 and Grothendieck's Standard Conjectures on p. 211.

set theoretically than are many other kinds of space such as differential manifolds or Riemann surfaces but we will skip all the details here.[8]

Any finite set of Diophantine equations in several variables defines a scheme, actually a special case called a *spectrum*, and general schemes are gotten by patching together compatible spectra just as a differential manifold patches together parts of n-dimensional real coordinate space $\mathbb{R}^n$. When a scheme X is presented as a topological space plus some algebraic structure then the points correspond to specialized forms of the equations for X. Notably, given integer polynomial equations the scheme organizes the corresponding equations on rational numbers and the corresponding equations modulo p for each prime number p. The algebra and topology of X capture all information about the equations including their solutions—in a form beautifully revealed by cohomology.[9]

The simplest useful cohomology for schemes relies on a notion of a *coherent sheaf of modules* $\mathcal{F}$ on a scheme X. Think of one sheaf $\mathcal{F}$ as posing one arithmetic problem all over X. The simplest problem is "choose a number," understanding that over some points of X "number" may mean a rational number, while over other points it will mean an integer modulo p for some prime number p depending on the point.

To stay with this simple, sketchy example for a moment, let $x \in X$ be some point where "number" means a rational number, and $y \in X$ a nearby point where "number" means an integer modulo 7. Notice that a rational number n/m has a well-defined integer value modulo 7 so long as the denominator m is not divisible by 7 when expressed in least terms. A *local solution* to our problem is a compatible selection of one "number" at each point in some region $U \subseteq X$, where a typical requirement of compatibility is this: if rational number n/m is chosen at x then the value of n/m modulo 7 must be chosen at y. We must choose "the same" solution at x and y, in this sense, although of course one is a rational number and the other an integer modulo 7.

If you think of any sheaf $\mathcal{F}$ as posing a problem all over X, then a *local section* of $\mathcal{F}$ is a choice of compatible solutions at each point in some region $U \subseteq X$, while a *global section* of $\mathcal{F}$ is one solution varying compatibly over all of X. In other words a global section is a local section for the special case where $U = X$ is the entire scheme. It is often comparatively easy to work on some problem locally in little regions, while the important question is to work on it globally.

Cohomology originally described the holes in topological spaces such as Riemann surfaces [Totaro, 2008, p. 389–91]. The first cohomology group $H^1(S)$ of a Riemann surface S counts the holes in S by measuring how

[8]Ellenberg [2008] gives a brief introduction. McLarty [2008] compares two set theoretic definitions of schemes, using functors or using topological spaces.

[9]Schemes were created to work with cohomology [McLarty, 2007].

much difference it makes to integrate a form on S along different routes. If the integral of a *holomorphic* 1-*form* α along one path in S differs from the integral of that same form along another path with the same endpoints, then those two paths together must surround at least one hole. Knowing how many different results you can get from integrating a single holomorphic 1-form α on S along different paths tells how many holes there are. This topological feature of S controls a great deal of complex analysis on S via deep theorems like the Riemann–Roch theorem.

The cohomology of a scheme X measures obstacles to passing from local to global solutions. Depending on the choice of a "problem" in the form of a sheaf $\mathcal{F}$ on X, there may be ways to patch together local solutions to $\mathcal{F}$ over any two little overlapping regions of X, which are compatible over any three little jointly overlapping regions, but which give different cumulative results when they travel around X by different routes—so they are not compatible all over X at once—sort of the way that integrating one form α along different paths between the same endpoints on a Riemann surface S can give different results. Such patching gives local solutions but not global. The first cohomology group $H^1(X, \mathcal{F})$ measures how many different ways this can happen in solving the problem $\mathcal{F}$ on the space X and so gives some measure of the "shape" of X in a way that expresses a great deal of arithmetic on X. The higher cohomology groups refine this.

As odd as that may sound it gives well organized access to arithmetic information in concrete cases. Wiles gets his access through coherent cohomology and other more intricate cohomology theories touched on below. The sheaves and cohomology groups of interest are consistently very small. They are continuum sized at most but contain quite complicated information. Even before going to curves or higher dimensional spaces, the cohomology of 0-dimensional single-point arithmetic schemes already includes the Galois theory of all algebraic number fields.

§4. Grothendieck's strategy.

> We will ignore any set theoretic difficulties. These can be overcome with standard arguments using universes. [Fantechi et al. 2005, p. 10]

Logicians complain about careless appeals to set theoretic power such as invoking the axiom of choice to show the field of rational numbers has an algebraic closure. It requires choice to show all fields have algebraic closures. Well, strictly speaking, relative to ZF, it *requires* the boolean prime ideal theorem which is weaker than the axiom of choice. Algebra textbooks rarely go into that subtlety. Anyway the whole issue is a gratuitous complication for countable fields like the rationals.

It is natural for a logician to suspect this is how large sets come into algebraic geometry. Perhaps the theory is framed to include arbitrarily large schemes that are not really of interest?[10] But no, that is not the reason. Even small schemes and sheaves bring in large sets because of a perspective that ought to interest logicians: Grothendieck's strategy for the vast complicated data of arithmetic was to create explicit organizing tools on a scale never seen before him—or, more accurately, never except in set theory organizing the entire universe of sets and in category theory. Even to work on a very small scheme, Grothendieck will place that scheme into artfully selected large contexts in ways we will describe. Not all number theorists like his perspective or even care to think about it. They use his and his school's theorems.

Grothendieck defined the cohomology of any sheaf of modules $\mathcal{F}$ over a scheme X not by the internal nuts and bolts of $\mathcal{F}$ but by the relations of $\mathcal{F}$ to all other sheaves over X. The nuts and bolts come in later, only when and as they are needed for particular calculations. In his own words he handled the "prodigious arsenal" of sheaves on X in terms of "its most obvious structure, which appears so to speak 'right in front of your face,' which is to say the structure of a 'category' " [Grothendieck, 1985, p. P38]. He did this already in [Grothendieck, 1957] which is one of the most widely cited papers in mathematics. Given one space X he took an array of categories of related spaces and sheaves on them as one simply and explicitly organized workspace guiding proofs about X.

He would "approach these categories from a 'naïve' point of view, as if we were dealing with sets" [Grothendieck and Dieudonné, 1971, p. 19]. The point was not to seek strong set theoretic axioms. To the contrary, Grothendieck aimed to preserve what he likes calling the "childish ... incorrigible naïveté" of his geometry [1985, p. P32]. But having worked over Bourbaki's set theory in draft, both Dieudonné and Grothendieck knew these are proper classes on naive ZFC foundations and they knew of Tarski's inaccessibles. So Grothendieck decided: "to avoid certain logical difficulties, we will accept the notion of a Universe, which is a set 'large enough' that the habitual operations of set theory do not go outside of it" [Grothendieck, 1971, p. 146].

§**5. First steps beyond ZFC.** No one would try to understand Wiles [1995] without mastering the standard graduate textbook [Hartshorne, 1977]. Hartshorne's central Chapter III spends 80 pages on the cohomology used to prove all the geometric results in the rest of the book. He assumes basic theorems of homological algebra not normally proven in graduate textbooks and, suitably to his purpose, he does not prove them either (p. 203). The only source he specifies for proofs is Freyd's book *Abelian Categories* which vaguely describes its own foundation as "a set theoretic language such as"

[10] Some wonder if this is what Hartshorne means by the "utmost generality" [1977, p. xiv]. In fact he means dropping the restriction to Noetherian rings, an issue little related to size. Noetherian rings can be any size and non-Noetherian can be of any infinite size.

Morse–Kelley set theory (MK), but goes beyond that as well in at least one case [Freyd, 1964, pp. 14 and 131]. Subsection 5.1 below compares the theories NGB, MK, and ZFC+U.

Following Grothendieck's strategy Hartshorne (p. 207) defines the infinite series of cohomology groups

$$H^0(X,\mathcal{F}), H^1(X,\mathcal{F}), \ldots, H^n(X,\mathcal{F}), \ldots$$

of a sheaf of modules $\mathcal{F}$ over X by the derived functors of the *global section functor*. The global section functor Γ goes from the category $\mathfrak{Mod}(X)$ of sheaves of modules on X, to the category $\mathfrak{Ab}$ of Abelian groups.

$$\mathfrak{Mod}(X) \xrightarrow{\quad\Gamma\quad} \mathfrak{Ab}$$

It takes any sheaf of modules $\mathcal{F}$ over X to the group $\Gamma(\mathcal{F})$ of its global sections.

Derived functors have several equivalent definitions which Hartshorne uses in combination for particular problems and for theoretical purposes. The most concise says a derived functor is a *universal δ-functor* [Hartshorne, 1977, p. 206]. The particulars are not as important for us here as the pattern:

1. A δ-functor T^* on $\mathfrak{Mod}(X)$ is an infinite series of ordinary functors $T^i\colon \mathfrak{Mod}(X) \to \mathfrak{Ab}$, $i \in \mathbb{N}$ plus natural transformations δ^i with a certain relation to exact sequences in $\mathfrak{Mod}(X)$.
2. A morphism $\eta^*\colon T^* \to S^*$ to another δ-functor S^* on $\mathfrak{Mod}(X)$ is a suitable infinite series of natural transformations $\eta^i\colon T^i \to S^i$.
3. A δ-functor U^* on $\mathfrak{Mod}(X)$ is *universal* if: for every δ-functor T^* every natural transformation $\eta^0\colon U^0 \to T^0$ of ordinary functors extends to exactly one δ-functor morphism $\eta^*\colon U^* \to T^*$.

In ZFC the categories $\mathfrak{Mod}(X)$ and $\mathfrak{Ab}$ and the functors between them are all proper classes. This definition quantifies over the functors. Probably everything in [Hartshorne, 1977] can be formalized in NGB although NGB puts limits on such familiar ideas as mathematical induction as we point out in Subsection 5.1 below. And apart from those limitations, formalizing this mathematics in NGB requires circumlocution around some natural collections of proper classes.

This characterization of a universal δ-functor transparently defines a category with δ-functors on $\mathfrak{Mod}(X)$ as objects and morphisms between them as arrows. In fact universal δ-functors are called *universal* because they are universal objects for a certain functor with this category of δ-functors as domain. That is the functor taking each δ-functor T^* to its zero part T^0. It is a natural way to think about the subject. But the category of δ-functors and morphisms is a superclass with each object and arrow a proper class. Such superclass categories are kept implicit throughout the textbook, never explicit. This is what I mean by circumlocution.

Superclass categories are explicit in the more advanced Hartshorne [1966]. This book defines δ-functoriality in terms of *derived categories* where each single morphism is a proper class.[11] Hartshorne quantifies over these superclass derived categories while keeping the set theory utterly unobtrusive as it should be.

The idea of a category of δ-functors is so obvious and unproblematic that it can safely be left implicit. Yet each category of δ-functors is a superclass of proper classes. Hartshorne's textbook reasonably elides such issues. The explicit language of his textbook goes only so far beyond ZFC as the conservative extension NGB. If universes one day become a less fraught subject then perhaps the more highly organized functorial tools used in research will become more accessible to students.

5.1. Comparing extensions of ZFC. Both NGB and MK extend ZFC by positing a class U containing all sets. Thus U is not itself a set; and with it they posit many subclasses of U which are also "too big" to be sets and so are called *proper classes*. The elements of proper classes are sets, and no proper class is an element of any collection in NGB or MK. The great difference is that NGB only allows quantification over sets in defining a class while MK can also quantify over classes to define a class.

Any model $\mathcal{M}$ of ZFC has a minimal extension to a model $\mathcal{M}'$ of NGB. Working outside of the model $\mathcal{M}$ we can form the collection $|\mathcal{M}'|$ of all subsets of the domain $|\mathcal{M}|$ definable with parameters in $\mathcal{M}$ in the language of ZFC. Of course each set α in $|\mathcal{M}|$ defines itself by the formula $x_1 \in \alpha$, so $|\mathcal{M}| \subsetneq |\mathcal{M}'|$. Form the model $\mathcal{M}'$ with this larger domain and the natural membership relation. Definable subsets which were not already sets in $\mathcal{M}$ become proper classes in $\mathcal{M}'$. Proper classes cannot enter into the specification of any set or class in NGB so $\mathcal{M}'$ is a model of NGB with no need to iterate this expansion. Relations among sets are unaltered by this process so any statement true in a model $\mathcal{M}$ of ZFC remains true of sets in the corresponding model $\mathcal{M}'$ of NGB.

Thus NGB is a conservative extension of ZFC and cannot prove consistency of ZFC. Mostowski [1950, p. 113] pinpoints the crucial fact: NGB can use the class U of all sets to formulate a truth predicate for Gödel numbers of formulas of ZFC but this truth predicate uses one (existentially) quantified class variable. So in NGB this predicate does not define a set or class of Gödel numbers of "true" formulas. Because formulas using this predicate do not define sets or classes in NGB, mathematical induction does not apply to them, so NGB can do little with this predicate and notably cannot use it to prove consistency of ZFC. The MK axioms allow impredicative definition of classes and so can use this truth predicate to prove consistency of ZFC.

[11] Carter [2008] gives a philosophic account of her work using the same *category of fractions* technique in a set theoretically smaller problem of geometry.

The extension ZFC+U is far stronger since it makes the universe U a set, which has a powerset $\mathcal{P}(U)$, which has a powerset $\mathcal{P}^2(U)$ in turn and so on through one higher rank $\mathcal{P}^\beta(U)$ for each ordinal number β, more commonly written as $V_\beta(U)$. By definition U models ZFC and it is straightforward to verify that the powerset $\mathcal{P}(U)$ models MK. The subsets of U with lower rank than U are in fact all elements of U and appear as sets in this model of MK, while the subsets with the same rank as U appear as proper classes.

§6. Grothendieck universes. We have seen that a standard textbook uses NGB and hints at superclasses of proper classes. Advanced results cited in applications use impredicative definitions of classes, which means MK rather than NGB. Other standard references use an impredicative theory of superclasses. This is still vastly weaker than ZFC+U but there is no assignable limit to how far it will go and there is really no reason to keep track of it. Anyone trying to minimize the logical assumptions of the number theory can use far less than ZFC in principle. The reason to go beyond ZFC is to give a safe and simple foundation for the published proofs.

With Grothendieck, we regard universes as a naïve way of treating all the categories of interest as sets.[12] We avoid the distinction of sets from proper classes at the heart of NGB and MK let alone the distinction of superclasses from proper classes required to go one rank higher. We avoid issues of definability which would be invoked to show when variables over superclasses can be replaced by explicit constructions on proper classes. Issues of definability may be extremely valuable for some problems but we can look just at the ones that are, when they are, rather than build certain ones into the foundations. Our foundation ZFC+U merely says there is a set U with a few natural closure properties.

Grothendieck's favorite stated reason for using universes is functor categories.[13] For categories $\boldsymbol{A}$ and $\boldsymbol{B}$ he will form the category $\boldsymbol{B}^{\boldsymbol{A}}$ of all functors from $\boldsymbol{A}$ to $\boldsymbol{B}$. We have seen how their use in EGA III lies behind a crucial step in Wiles [1995, pp. 486–7]. Any one use of this will be only one or a few ranks higher set theoretically than $\boldsymbol{A}$ and $\boldsymbol{B}$, depending on exactly how we define categories in terms of sets. But then we might need yet higher rank for another functor category $\boldsymbol{C}^{(\boldsymbol{B}^{\boldsymbol{A}})}$. Grothendieck prefers to treat all these as sets. A universe U lets him do this. So long as the initial categories are no bigger than U (and so they are, from the viewpoint of ZFC, no bigger

[12]See Lurie [2009, pp. 50f.] for a recent mathematical discussion adopting universes, and noting that as many universes exist as there are strongly inaccessible cardinals. Lurie follows the common practice of assuming just "enough" successively larger universes without trying to keep track of how many he wants.

[13]Grothendieck [1971, pp. 146f.] dryly promises a definitive treatment by Claude Chevalley and Pierre Gabriel in the year 2000. Until then he offers his own [1957] noting it is very incomplete even for his purposes.

than proper classes) then all the finitely iterated functor categories will be sets in ZFC+U.

At the interface of the original arithmetic application and the general theory lie set theoretically large *sites*. The cohomology of topological spaces handles any topological space T in terms of the set $\mathrm{Ouv}(T)$ of all open subsets of T. These form the *site* for that cohomology. They form a set no bigger than the powerset of the set of points of T. Wiles [1995] works primarily in that framework, as does Hartshorne [1977] as described in Sections 3–5 above. But Mumford and Tate [1978] give a beautifully concise account of how the original arithmetic application used the *étale cohomology* of a scheme X which replaces open subsets of X by *étale maps* $X' \to X$. These collectively form the *petit étale* site for *X*. The word "petit" here refers to an algebro-geometric distinction between *gros* and *petit* étale sites. It is not based on set theoretic size. From the viewpoint of ZFC the petit étale site of a scheme is not a set but a proper class. Grothendieck points to several technical tricks to avoid these proper classes and to the inconvenience of these tricks in practice and affirms this set theoretically large site as the right one for étale cohomology (SGA 4, p. 307).

The use of set theoretically large sites replaces many of the sets described in Sections 3–5 with proper classes and thus replaces most of the proper classes in those sections with superclasses. It replaces all the superclasses we have talked about with whatever you like to call collections of superclasses three ranks above ZFC. Once again, there is no need for any of this if the goal is to find the weakest logic sufficient for the arithmetic proofs in principle. We have the different goal of formalizing the proofs as they are published. Once you go this far above ZFC there is no reason to stop short of simply using universes.

§7. Deligne and SGA $4\frac{1}{2}$. Deligne [1977] provides the kind of expert introduction to étale cohomology that too few advanced techniques ever get. Oddly, some people suppose this book makes Deligne's proof of the last Weil conjecture independent of universes and of the other SGA.

The book explicitly uses set theoretically large sites [1977, p. 23] and implicitly uses universes in other ways. The goal is to be "clearer than SGA 4 … but not claim to give a complete proof" (p. 2). For proofs of such necessary steps as Poincaré duality and the trace formula Deligne cites his own articles in SGA 4 which explicitly use universes. This book is meant to give an adequate working background for Deligne's proof of the last Weil conjecture based on no more than some cohomology of topological spaces plus "un peu de foi" (p. 1). But Deligne never suggests faith should replace proofs in the end.

While Deligne often uses universes he stresses in conversation that they are a convenience technically eliminable in favor of ZFC. The general theorems used in practice can always be given in terms of individual sheaves on small sites (where "small" means provably existing in ZFC), without ever looking at whole categories of sheaves let alone categories of categories of them and so on. This is a recipe for eliminating universes from any use of Grothendieck's cohomology in number theory or anywhere else. Though it is obvious in practice that it could always be done, it is not done in publications, and it has never been made a precise metatheorem. Anyone interested in that should give it a try.

By such means the great cohomological proofs like Deligne [1974], or Faltings [1983], or Wiles [1995] never need to go beyond ZFC. But in fact these three as written and published all use Grothendieck's tools. All three either cite proofs in EGA and SGA using universes or cite sources that do.

The point is that mathematics is not only technical. Deligne [1998] explains the practical value of Grothendieck's high level organization, notably the value of toposes. He explains that Grothendieck would not describe single structures but would describe a category forming a workspace around each one. Grothendieck did not only define a scheme for each very small geometrically or arithmetically plausible commutative ring, but for all commutative rings:

> If the decision to let every commutative ring define a scheme gives standing to bizarre *schemes*, allowing it gives a *category of* schemes with nice properties. [Deligne, 1998, p. 13].

That category is the easy, natural way to work with schemes. And Grothendieck would not only work with very small geometrically or arithmetically plausible sheaves on a given scheme, but with the topos of all sheaves on it, because this led to the right, natural definition of cohomology as a δ-functor:

> Grothendieck had shown that, given a category of sheaves,[14] a notion of cohomology groups results. [Deligne, 1998, p. 16].

In a recent talk Deligne makes the same point about motives: Grothendieck does not seek to define motives piecemeal by their nuts and bolts, but by their intrinsic relations in a category of motives. See [Deligne, 2009, minute 5].

This is the strategy that produced modern cohomological number theory. The goal is not at all to posit large categories. The goal is to posit adequate categories and treat them as sets. The only simple, conceptual way to do that which is yet known uses universes—and it is eliminable in principle at the cost of complicating the work.

7.1. One possible strategy for a metatheorem. Perhaps the overall nature of Grothendieck's cohomology theory could be retained within ZFC by replacing universes with "small universes." Those are sets V_β for limit ordinals β which provably exist in ZFC. They model all of ZFC except the

[14]I.e., a Grothendieck topos.

replacement axiom scheme. But in order to work, the limit ordinal β must be large enough to bound all the required transfinite inductions, notably in the proof that certain categories have "enough injectives" [Grothendieck, 1957]. If it can be proved in ZFC that for any given site some limit ordinal β suffices to bound all the inductions needed for the cohomology of that site then by replacement there would also be suitable β for any set of sites. A single proof in cohomological number theory never uses more than a set of sites. So if that difficulty can be overcome then each of these proofs could be given within ZFC with no perceptible damage to the theoretical organization of the proof.

§8. Functoriality and weak proofs. Everyone would like to lighten all proofs in number theory (or any mathematics) as much as they can in any way that they can. For many number theorists that would include eliminating functorial tools. All number theorists share Lenstra's goal of solving equations while many do not yet share his amusement:

> Hendrik Lenstra, in his lecture to the conference, recounted that twenty years ago he was firm in his conviction that he DID want to solve Diophantine equations, and that he DID NOT not wish to represent functors—and now he is amused to discover himself representing functors in order to solve Diophantine equations! [Mazur, 1997, p. 245, emphasis in the original].

Funny things can be true. The evidence is that functors make arithmetic easier. Indeed, up to this time, they make Wiles's proof feasible.

A central functorial tool in Wiles's proof and a great deal of other number theory is *group cohomology*.[15] This assigns to each group G, and Abelian group A acted on by G, an infinite series of cohomology groups

$$H^0(G, A), H^1(G, A), H^2(G, A), \dots$$

Washington [1997, p. 103] explains how Wiles's proof uses primarily these first three terms H^0–H^2, and he describes their concrete arithmetic meaning. At the same time he explains how these groups appear as values of functors H^0, H^1, H^2 in an infinite series of functors H^n with a strong analogy to cohomology in topology.

No one who has looked at the proof doubts that this cohomology and all the rest of Wiles [1995] could be routinely unwound, doing some currently unknown amount of damage to the theoretical organization, to work in sets of rank perhaps 7 or 8 over the natural numbers. This would not require solving any new arithmetic problems but only eliminating some currently unknown large number of general definitions (which, of course, occur nested in each other) in cohomology, in favor of their specific applications in

[15]For the origin of this cohomology see [Mac Lane, 1988] and for more historical detail Basbois [2009].

arithmetic. This routine elimination is far weaker than Macintyre's program because it uses far stronger logic. It assumes full comprehension in each rank and no restriction on the formulas used in inductions.

Macintyre makes a far stronger claim about the cohomology when he says "there is no evidence at all that Base Change or the Trace Formula has any essentially higher-order content" [forthcoming, MS pp. 14–15]. Each single use of these theorems represents rather explicit calculations in arithmetic. These uses very likely can be either unwound into PA itself, or else justified in a higher-order conservative extension of PA which does restrict induction and comprehension perhaps along the lines of [Takeuti, 1978]. That would complete Macintyre's argument that FLT has nothing to do with the Gödel phenomenon—i.e., it has nothing to do with any facts of arithmetic that actually require axioms stronger than PA.

Macintyre presents evidence but also shows how his claim remains to be verified by a great deal of further work in arithmetic. This work may be very enlightening and probably will not be easy. Quite aside from the general theorems using universes, the most concrete appeals to group cohomology are on their face several ranks over the natural numbers. The concrete version of H^1 is a group of equivalence classes of crossed homomorphisms from a Galois group to its number field. Getting the requisite facts on these groups into PA or a higher-order conservative extension of PA will be no routine exercise. It will require serious new arithmetic.

§9. Foundations.

> The point of foundations is not to arbitrarily restrict inquiry but to provide a framework wherein one can legitimately perform those constructions and operations that are mathematically interesting and useful.
>
> —Herrlich and Strecker [1973, p. 331].

> The really interesting foundational matter is finding genuine unremovability of Universes in the integers. In fact, there is presently no genuine unremovability of Universes for any statement about sets of limited rank! This is because, e.g., regularity properties about projective sets of real numbers either can be proved in ZFC or require large cardinals far beyond Universes.
>
> —Friedman post Apr 8, 1999 on FOM.[16]

Much of the large apparatus of Wiles [1995] will one day be by-passed in favor of more direct use of PA. It will not all be easy, and it is impossible to know now how far it will go. At the same time progress will continue making the functorial apparatus swifter and more accessible—in Grothendieck's

[16]E-mail list FOM, Friedman post titled "`Using Universes? The expert speaks again, Thu Apr 8, 1999`" archived at `cs.nyu.edu/pipermail/fom`.

terms, more "naïve." This will not all be easy, and it is impossible to know now how far it will go. Both those projects will go on, as they are going on, apart from what any logician thinks of either one. And both will advance arithmetic.

For now, though, we must look to high levels of organization just as Wiles did because he wanted to finish his proof. We are led to EGA III and SGA 1 and SGA 4, as Wiles's sources are. We arrive at universes.

We want more than one thing one from foundations. We study what logic is legitimate in mathematics. We seek to confirm or refute the genuine unremovability of various strong set theoretic axioms for statements about sets of limited rank. Without defending Friedman's sweeping claim about this I entirely agree that cohomological number theory offers no such unremovability of universes. I doubt anyone interested in the subject ever thought it would. Yet this number theory presents a large body of legitimate, interesting and useful constructions and operations using universes—if we can agree that everything Wiles [1995] uses in fact is legitimate, interesting, and useful.

REFERENCES

[1972] M. Artin, A. Grothendieck, and J.-L. Verdier, *Théorie des topos et cohomologie etale des schémas*, ***Séminaire de Géométrie Algébrique du Bois-Marie, 4***, Springer-Verlag, 1972, three volumes, generally cited as SGA 4.

[2003] J. Avigad, *Number theory and elementary arithmetic*, ***Philosophia Mathematica***, vol. 11 (2003), pp. 257–284.

[2009] N. Basbois, ***La naissance de la cohomologie des groupes***, Ph.D. thesis, Université de Nice Sophia-Antipolis, 2009.

[2008] J. Carter, *Categories for the working mathematician: Making the impossible possible*, ***Synthese***, vol. 162 (2008), pp. 1–13.

[1997] G. Cornell, J. Silverman, and G. Stevens (editors), ***Modular forms and Fermat's Last Theorem***, Springer-Verlag, 1997.

[1974] P. Deligne, *La conjecture de Weil I*, ***Publications Mathématiques. Institut de Hautes Études Scientifiques***, (1974), no. 43, pp. 273–307.

[1977] P. Deligne (editor), *Cohomologie étale*, 1977, ***Séminaire de Géométrie Algébrique du Bois-Marie; SGA 4 1/2***, Springer-Verlag. Generally cited as SGA 4 1/2, this is not strictly a report on Grothendieck's Seminar.

[1998] ———, *Quelques idées maîtresses de l'œuvre de A. Grothendieck*, ***Matériaux pour l'histoire des mathématiques au XX^e siècle*** (***Nice, 1996***), Société Mathématique de France, 1998, pp. 11–19.

[2009] ———, *Colloque Grothendieck, Pierre Deligne*, video by IHES Science, on-line at `www.dailymotion.com/us`, 2009.

[1973] P. Deligne and M. Rapoport, *Les schémas de modules de courbes elliptiques*, ***Modular functions of one variable, II***, Lecture Notes in Mathematics, vol. 349, Springer-Verlag, New York, 1973, pp. 143–316.

[2008] J. Ellenberg, *Arithmetic geometry*, ***Princeton companion to mathematics*** (T. Gowers, J. Barrow-Green, and I. Leader, editors), Princeton University Press, 2008, pp. 372–383.

[1983] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, ***Inventiones Mathematicae***, vol. 73 (1983), pp. 349–366.

[2005] B. Fantechi, A. Vistoli, L. Gottsche, S. L. Kleiman, L. Illusie, and N. Nitsure, ***Fundamental algebraic geometry: Grothendieck's FGA explained***, Mathematical Surveys and Monographs, vol. 123, American Mathematical Society, Providence, 2005.

[1964] P. Freyd, ***Abelian categories: An introduction to the theory of functors***, Harper and Row, 1964, reprinted with author commentary in: ***Reprints in Theory and Applications of Categories***, (2003), no. 3, pp. −25–164, available on-line at `/www.emis.de/journals/TAC/reprints/articles/3/tr3abs.html`.

[1957] A. Grothendieck, *Sur quelques points d'algèbre homologique*, ***Tôhoku Mathematical Journal***, vol. 9 (1957), pp. 119–221.

[1971] ———, *Revêtements étales et groupe fondamental*, ***Séminaire de Géométrie Algébrique du Bois-Marie, 1***, Springer-Verlag, 1971, generally cited as SGA1.

[1985] ———, ***Récoltes et semailles***, Université des Sciences et Techniques du Languedoc, Montpellier, 1985, published in several successive volumes.

[1961] A. Grothendieck and J. Dieudonné, *Éléments de géométrie algébrique III: Étude cohomologique des faisceaux cohérents*, ***Publications Mathématiques. Institut des Hautes Études Scientifiques, Paris***, (1961), no. 11.

[1971] ———, ***Éléments de géométrie algébrique I***, Springer-Verlag, 1971.

[1966] R. Hartshorne, ***Residues and duality, lecture notes of a seminar on the work of A. Grothendieck given at Harvard 1963–64***, Lecture Notes in Mathematics, no. 20, Springer-Verlag, New York, 1966.

[1977] ———, ***Algebraic geometry***, Springer-Verlag, 1977.

[1973] H. Herrlich and G. Strecker, ***Category theory***, Allyn and Bacon, Boston, 1973.

[2009] J. Lipman and M. Hashimoto, ***Foundations of Grothendieck duality for diagrams of schemes***, Springer-Verlag, 2009.

[2009] J. Lurie, ***Higher topos theory***, Annals of Mathematics Studies, no. 170, Princeton University Press, Princeton, 2009.

[1988] S. Mac Lane, *Group extensions for* 45 *years*, ***Mathematical Intelligencer***, vol. 10 (1988), no. 2, pp. 29–35.

[2003] A. Macintyre, *Model theory: Geometrical and set-theoretic aspects and prospects*, this Bulletin, vol. 9 (2003), no. 2, pp. 197–212.

[forthcoming] ———, *The impact of Gödel's incompleteness theorems on mathematics*, ***Horizons of truth: Proceedings of Gödel centenary, Vienna, 2006***, forthcoming.

[1977] B. Mazur, *Modular curves and the Eisenstein ideal*, ***Publications Mathématiques. Institut des Hautes Études Scientifiques***, vol. 47 (1977), pp. 133–186.

[1997] ———, *Introduction to the deformation theory of Galois representations*, ***Modular forms and Fermat's Last Theorem*** (G. Cornell, J. Silverman, and S. Stevens, editors), Springer-Verlag, 1997, pp. 243–312.

[2007] C. McLarty, *The rising sea: Grothendieck on simplicity and generality I*, ***Episodes in the history of recent algebra*** (J. Gray and K. Parshall, editors), American Mathematical Society, 2007, pp. 301–326.

[2008] ———, *"There is no ontology here": visual and structural geometry in arithmetic*, ***The philosophy of mathematical practice*** (P. Mancosu, editor), Oxford University Press, 2008, pp. 370–406.

[1950] A. Mostowski, *Some impredicative definitions in the axiomatic set theory*, ***Fundamenta Mathematicae***, vol. 37 (1950), pp. 111–124.

[1978] D. Mumford and J. Tate, *Fields Medals IV. An instinct for the key idea*, ***Science***, vol. 202 (1978), pp. 737–739.

[2008] B. Osserman, *The Weil conjectures*, ***Princeton companion to mathematics*** (T. Gowers, J. Barrow-Green, and I. Leader, editors), Princeton University Press, 2008, pp. 729–732.

[1978] G. Takeuti, *A conservative extension of Peano Arithmetic*, ***Two applications of logic to mathematics***, Princeton University Press, 1978, pp. 77–135.

[2008] B. Totaro, *Algebraic topology*, ***Princeton companion to mathematics*** (T. Gowers, J. Barrow-Green, and I. Leader, editors), Princeton University Press, 2008, pp. 383–396.

[1997] L. Washington, *Galois cohomology*, ***Modular forms and Fermat's Last Theorem*** (G. Cornell, J. Silverman, and S. Stevens, editors), Springer-Verlag, 1997, pp. 101–120.

[1995] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem*, ***Annals of Mathematics***, vol. 141 (1995), pp. 443–551.

DEPARTMENT OF PHILOSOPHY
CASE WESTERN RESERVE UNIVERSITY
CLEVELAND, OH 44106, USA
E-mail: colin.mclarty@case.edu