Symmetric comprehension revisited

M. Randall Holmes

Version of 6/16/2017, 3 pm Boise time

1 Introduction

In our paper [**?**], we defined a rather baroque theory of superclasses, classes, and sets in which the criterion for sethood of a superclass was a symmetry criterion, the sets of any model of which would satisfy Quine's New Foundations (the theory described in [**?**]). We will not give the details of this here: we will give a simpler implementation of the same general idea, a predicative theory of classes and sets in which the criterion for a class to be a set is stated in terms of symmetry, and the sets of any model of which will similarly satisfy NF. This is not a contribution toward the consistency question for NF: we do not address the question of how to build a model of either of these class theories.

This work was motivated by a theorem of Forster and others ([**?**], pp. ???) to the effect that the stratified formulas of the language of set theory admit a semantic characterization as precisely the formulas which are invariant under setlike permutations in all models. This suggested to us that it ought to be possible to develop a semantic motivation for stratified comprehension (the comprehension of NF which is often dismissed as a "syntactical trick").

2 Models of the Predicative Theory of Classes with Pairing

Our metatheory will be the usual set theory ZFC: we will not use very much of it. The symbol $\in$ will represent the membership of the metatheory, and $\{x : \phi\}$ will represent the unique A such that $x \in A \leftrightarrow \phi$, for any formula of the language of the metatheory (which will be taken to include the language of our theories as a sublanguage). $\{x, y\}$ will represent $\{z : z = x \vee z = y\}$ ($\{x\}$ abbreviating $\{x, x\}$) and (x, y) will represent the usual ordered pair $\{\{x\}, \{x, y\}\}$. We will be systematic about providing notational distinctions between the basic operations of the metatheory (the set theory we are using) and the basic operations of the predicative theory of sets and classes that we are talking about.

Definition (model of predicative theory of sets and classes): A pair (C, ϵ) will be termed a model of the predicative theory of sets and classes (or, more briefly, a model of predicative class theory) iff C is a set, $\epsilon \subseteq C \times C$, and the following additional conditions hold (where $x \epsilon y$ means $(x, y) \in \epsilon$):

definition of internal sethood: Define S as

$$\{x \in C : (\exists y \in C : x \epsilon y)\}.$$

The elements of S are naturally understood as the sets of the model (C, ϵ) and we will call them "internal sets".

extensionality: We stipulate as part of the definition of a model of predicate class theory that $(\forall xy \in C : x = y \leftrightarrow (\forall z \in S : z \epsilon x \leftrightarrow z \epsilon y))$

An alternative stipulation is weak extensionality:

$$(\forall xyz \in C : z \in x \wedge (\forall u : u \in x \leftrightarrow u \in y) \rightarrow x = y).$$

If weak extensionality is used, we provide as a further stipulation that there is $0 \in S$ such that $(\forall x \in C : \neg x \epsilon 0)$.

comprehension: We will refer to the first-order language whose primitive predicates are ϵ and $=$ and whose quantifiers are understood to be bounded in C as the internal language of our predicative

theory of classes. We allow ourselves to extend this language by suitable definitions as we work, and we regard the internal language as a subset of our metalanguage with the qualification that quantifiers of the internal language must be explicitly bounded in C when translated to the metalanguage. We define $(\forall x \in S : \phi)$ as $(\forall x : (\exists y : x \epsilon y) \rightarrow \phi))$ and $(\exists x \in S : \phi)$ as $(\exists x : (\exists y : x \epsilon y) \wedge \phi))$, showing that we can bound quantifiers to the internal sets in the internal language.

We stipulate as part of the definition of a model of predicate class theory that for each formula ϕ in the first order language whose primitive predicates are ϵ and equality and whose quantifiers are understood to be restricted to S,

$$(\exists A \in C : (\forall x \in C : x \epsilon A \leftrightarrow x \in S \wedge \phi)).$$

The formula ϕ may contain free variables (parameters) understood to represent elements of C.

extension: For any $x \in C$, define $E(x)$ as $\{y \in S : y \epsilon x\}$. The extensionality property of the model tells us that E is a bijection, so for any set $X \in C$, it may be the case that there is an element $E^{-1}(X)$ of C (there is at most one such object for each subset X of C, but there may in fact be no such object). We will abbreviate $E^{-1}(\{x : \phi\})$ as $[x \mid \phi]$ [if we use weak extensionality, we stipulate that if $\{x : \phi\}$ is empty, $[x \mid \phi] = 0$, as uniqueness of the preimage under E fails in this case], so for example $[x|x = x]$ is the unique element v of C such that for any $y \in S$, $y \epsilon v$ (the universal class of the model (which contains all sets, not all classes)).

existence of sets and pairs: Define $\langle x, y \rangle$ as $[z|z = x \vee z = y]$ for any $x, y \in S$. We further stipulate as part of the definition of a model of predicative class theory that S is nonempty and that for each $x, y \in S$, $\langle x, y \rangle \in S$: this amounts to asserting that there is at least one element and adding the axiom of pairing for sets to our predicative theory of classes. Define $[x, y]$ as $\langle \langle x, x \rangle, \langle x, y \rangle \rangle$. Refer to objects $[x, y]$ as "internal (ordered) pairs". We note that we use (x, y) to represent the usual ordered pair in the metatheory. We may use the abbreviation $\langle x \rangle$ for $\langle x, x \rangle$.

definition of internalized relations and functions: An internal relation is an element of C whose extension is a collection of internal

pairs. The relational extension of an internal relation R is defined as $\{(x, y) : [x, y]\epsilon R\}$. An internal function is an internal relation whose relational extension is a function in the sense of the metatheory (this is a convenient way to say this, but note that we can express the fact that f is an internal function in the language of predicative class theory). An internal permutation is an internal relation whose relational extension is a bijection from S onto S (again, this is straightforward to express in the internal language of the model). For any internal function F and element x of S define $F[x]$ as the unique y such that $[x, y]\epsilon F$. We say that an internal function f fixes $s \in S$ iff $f[s] = s$.

completion: We have completed the definition of a model of the predicative theory of sets and classes at this point, and stated subordinate definitions which will be used in the following development.

3 Notions of Symmetry

In the following discussion, we work in a fixed model (C, ϵ) of the predicative theory of sets and classes.

definition of the j operation on internal permutations: If π is an internal permutation and $x \in C$, define $\pi\{x\}$ as $[v \mid (\exists u : u \epsilon x \wedge [u, v] \epsilon \pi]$.

If weak extensionality is used, define $\pi\{x\}$ as above only if the extension of x under ϵ is nonempty: otherwise define $\pi\{x\}$ as x.

Define $j(\pi)$ as $[z \mid (\exists x : z = [x, \pi\{x\}]]$, which we may write as $[[x, \pi\{x\}] \mid x = x]$. Note that if $\pi\{x\} \notin S$, $[x, \pi\{x\}]$ does not exist.

Notice that $\pi\{x\}$ and $j(\pi)$ are elements of C (objects of the class theory we are talking about) while j is a function in the sense of the metatheory.

It is clear from the way we have defined these notions that $\pi\{x\}$ will belong to C if x belongs to C, but it is important to notice that $\pi\{x\}$ will not necessarily belong to S if x belongs to S. $j(\pi)[x]$ is only defined for $x \in S$ of course, but may fail be defined for some elements of S: if $\pi\{x\} \in C \setminus S$, $j(\pi)[x]$ will not be defined.

If π is an internal permutation, define $\pi^{\leftarrow}$ as $[[y, x] \mid [x, y] \epsilon \pi]$. Clearly this will also be an internal permutation.

Observe further that by extensionality $j(\pi)[x] = j(\pi)[y]$ implies $x = y$, and further, that if both $j(\pi)[x]$ and $j(\pi^{\leftarrow})[x]$ are defined for all $x \in S$, $j(\pi)$ will itself be an internal permutation. We define $j^0(\pi)$ as π and $j^{n+1}(\pi)$ as $j(j^n(\pi))$ (when $j^n(\pi)$ itself is an internal permutation; if $j^n(\pi)$ is not an internal permutation, $j^{n+1}(\pi)$ is not defined). Each j^n is a function of the metatheory.

setlike permutation: We say that an internal permutation π is setlike iff $j^n(\pi)$ is defined for every n. We say that an internal permutation π is n-setlike iff $j^n(\pi)$ is defined. It should be noted that the assertion that π is setlike cannot be expressed in the language of the predicative theory of sets and classes (at least, not in any obvious way), but the assertion that π is n-setlike can be so expressed for each n.

n-symmetry: We say that $x \in C$ is n-symmetric iff $j^{n-1}(\pi)\{x\} = x$ for every $(n-1)$-setlike internal permutation π. We say that $x \in C$ is n-

symmetric with support $A \in S$ iff $j^{n-1}(\pi)\{x\} = x$ for all $(n-1)$-setlike π for which $j^{n-1}(\pi)[A] = A$.

4 The Main Theorem

Definition (stratified formula, the theories NF(U)): A formula ϕ in the language of equality and membership is said to be *stratified* iff there is a function σ from variables appearing in ϕ to natural numbers such that for each subformula $x = y$ of ϕ we have $\sigma(x) = \sigma(y)$ and for each subformula $x \in y$ of ϕ we have $\sigma(x) + 1 = \sigma(y)$. The function σ is called a *stratification of* ϕ.

The axioms of NF are extensionality and the scheme of stratified comprehension ("$\{x : \phi\}$ exists", for each stratified formula ϕ); NFU is obtained if the axiom of extensionality is replaced with the weak axiom of extensionality.

For the original definition of this theory, see [**?**]; for an extended discussion of these theories in modern terms, see [**?**]. The consistency of NF is an open question; the consistency of NFU was established by Jensen in [**?**].

Theorem: If a model (C, ϵ) of predicative class theory satisfies the condition that for each $x \in C$, x belongs to S iff x is 3-symmetric with a support then the structure (S, ϵ) is a model of Quine's NF [or of Jensen's NFU if weak extensionality is used], with the relation ϵ implementing the membership relation of NF(U). Note the change in the domain of the structure to S rather than C. Notice that the condition that each element of C is an internal set iff it is 3-symmetric with a support can be expressed in the internal language of the predicative class theory, so this condition does correspond to a proposed set comprehension axiom to be adjoined to that theory (a description of such a development is given in an appendix).

Proof of theorem: To show this, it is necessary and sufficient to show that for any stratified formula ϕ of the internal language of predicative class theory [in this context, the conditions imposed on $=$ and $\in$ in the definition of stratification above are imposed on $=$ and ϵ respectively] in which quantifiers are bounded in S and in which each free variable is understood to denote an object which is 3-symmetric with a support, $[x \mid \phi]$ will be 3-symmetric with a support. We provided in the definition of a model of predicative class theory that there is at least one

element so that an arbitrary element can serve as support for a class (such as $[x \mid x \neq x]$) which is 3-symmetric with no need for a support.

We first argue that for any internal permutation f, if f is 2-setlike, so is $j(f)$. That f is 2-setlike means that $j^2(f)[x]$ is defined for every $x \in S$. Our aim is to show that $j(f)$ is 2-setlike, that is, that $j^3(f)[x] \in S$ for every $x \in S$. Under our hypothesis, $x \in S$ holds iff there is s such that for any 2-setlike g such that $j^2(g)[s] = s$, we have $j^2(g)\{x\} = x$. We argue that $j^2(f)\{x\} \in S$ (and so is equal to $j^3(f)[x]$ which is thus defined) by showing that it is 3-symmetric with support $j^2(f)[s]$. Suppose that g is a 2-setlike permutation such that $j^2(g)[j^2(f)[s]] = j^2(f)[s]$. Define $(f \cdot g)$ for f, g internal functions as the internal function h such that $h[x] = f[g[x]]$ for all $x \in S$ (which clearly exists). It follows that $(j^2(f)^{\leftarrow} \cdot j^2(g) \cdot j^2(f))$ fixes s, and this is the same as $j^2(f^{\leftarrow} \cdot g \cdot f)$ fixing s, from which it follows by symmetry of x that $j^2(f^{\leftarrow} \cdot g \cdot f)\{x\} = x$, from which it follows that $j^2(g \cdot f)\{x\} = j^2(f)\{x\}$, from which it follows that $j^2(g)\{j^2(f)\{x\}\} = j^2(f)\{x\}$, which is exactly what we need to show to show that $j^2(f)\{x\}$ is an internal set with support $j^2(f)[s]$, from which we observe that every $j^2(f)\{x\}$ is an internal set, so equal to $j^3(f)[x]$, so $j(f)$ is 2-setlike. One should also note that the same argument applies to the internal inverse of f. This further implies that any 2-setlike internal permutation is n-setlike for each natural number n (note that this last assertion cannot be formulated in any obvious way in the internal language of the model, though each of its instances for specific n can be so formulated).

We observe that for any internal permutation f, $x \epsilon y \leftrightarrow f[x] \epsilon j(f)[y]$ and so for any k, $x \epsilon y \leftrightarrow j^k(f)[x] \epsilon j^{k+1}(f)(y)$. Of course $x = y \leftrightarrow j^k(f)[x] = j^k(f)[y]$. Note that if f is 2-setlike we can carry out these transformations for any k because of the result of the previous paragraph.

Now observe that we can transform any stratified formula ϕ into a form where each variable x which is assigned type i in a given stratification of ϕ is replaced with $j^i(f)[x]$, without changing the truth value of the formula. From each quantified variable we can drop this decoration, because the relational extensions of the internal permutations $j^k(f)$ are permutations of the domain S of the quantifiers. We can then see that any $[x \mid \phi]$ is fixed by $j^{k+1}(f)$ for any f if k is the type assigned to x in a stratification of the formula ϕ and $j^i(f)$ fixes each parameter

a in ϕ which is assigned type i in the stratification of ϕ that we use. We can arrange for k to be 2: we can always raise k, by uniformly raising all types in a stratification, and it is a well-known result of Grishin (in [**?**]) that the axioms of NF using types 0,1,2,3 give a full axiomatization of NF, and any abstract $\{x : \phi\}$ provided by such an axiom has a stratification with the type of x less than or equal to 2. All of this applies equally well to the NFU case if weak extensonality is used.

It remains to show that we can construct a single s such that each condition $j^i(f)[a] = a$ determined by a parameter a assigned type i in the stratification of ϕ will hold if $j^2(f)[s] = s$ holds: we need devices to merge multiple support elements into one and revise their types. If $i < 2$, then $j^2(f)[\iota^{2-i}(a)] = \iota^{2-i}(a)$ is equivalent to $j^i[a] = a$, where $\iota(a) = \langle a \rangle$. If $i > 2$, we need to follow a different tack. Notice that if $j^{2+i}(f)[a] = a$ is desired, then $j^2(j^{i-1}(f))\{a\} = a$ is desired, for which it is sufficient for $j^{2+i-1}(f)[s] = s$ to hold, where s is a support of a. But this can be iterated: pass to a support of s, then a support of a support of s, and so forth, until the exponent is lowered to 2.

Finally, to enforce any finite number of conditions $j^2(f)[a_i] = a_i$ at once, we show how to collapse two conditions into one. Suppose a and b are internal sets. If f is a setlike internal permutation such that $j^2(f)$ fixes a support of a support of $[a, b]$ (which is an internal set by the axiom of pairing so has a support, as does each of its supports), then $j^6(f)$ fixes $[a, b] = \langle \langle a, a \rangle , \langle a, b \rangle \rangle$ whence $j^2(f)$ fixes a and $j^2(f)$ fixes b. If no conditions at all need to be enforced, use any internal set as the support.

It is possible to give an explicit description of supports of unordered pairs (making it possible to see concretely why we can merge supports). If we are using weak extensionality, observe that we can without loss of generality suppose that a support witnessing 3-symmetry is a set of sets (an element of S whose extension is a subset of S): if a support is an atom, it can be replaced with 0, the internal set whose extension is empty, and if a support is an internal set whose extension contains atoms, it can be replaced by the internal set whose extension is obtained by dropping all atoms from the original extension: in both cases, a map $j^2(\pi)$ fixes the original and the modified supports under exactly the same circumstances. Let $x + y$ denote $[z : z \epsilon x \vee z \epsilon y]$. Let $x - y$

denote $[z \mid z\epsilon x \wedge \neg z\epsilon y]$. Let $\langle x, y, z\rangle$ represent $\langle x, y\rangle + \langle z\rangle$. Let $a, b \in S$ be sets of sets in the sense indicated above. Let c, d, e, f, g, h be any six distinct elements of S. Then if any $j^2(F)$ fixes

$$s = \langle\langle c\rangle\rangle + \langle\langle c, d\rangle\rangle + \langle\langle d, e\rangle\rangle + \langle\langle e, f\rangle\rangle + \langle\langle f, g\rangle\rangle + \langle\langle g, h\rangle\rangle$$

$$+[x + \langle c, d, e\rangle - \langle f, g, h\rangle \mid x\epsilon a] + [x + \langle f, g, h\rangle - \langle c, d, e\rangle \mid x\epsilon b],$$

then $j^2(F)$ fixes both a and b. Hints: note that the objects c, d, e, f, g, h must be fixed by F if this object is fixed by $j^2(F)$, then observe that under these conditions the action of $j(F)$ on $[x+\langle c, d, e\rangle-\langle f, g, h\rangle \mid x\epsilon a]$ gives full information about the action of $j(F)$ on a, and that the three different kinds of elements of the extension of s (ones that identify the six special objects, ones that partially indicate elements of a, and ones that partially indicate elements of b) can be distinguished by cardinality in the first case and by which special objects are in their extension in the second case.

5 Conclusion and Strange Further Results

Why is this result interesting? It has often been said that NF lacks motivation because the restriction on comprehension used is based on a syntactical trick. Here we give a criterion for sethood in a class theory which is of a semantic nature, relying on actual set theoretical properties of the classes to be picked out as sets. As is well known, ordinary set theory can be motivated as an extension of predicative class theory by the criterion that classes are sets iff they are **small**. Here we suggest the alternative criterion that a class is a set if it is **symmetric** in a certain precise sense. And this semantic criterion for sethood in the context of predicative class theory entails the purportedly semantically unmotivated comprehension scheme of NF(U).

We have a philosophical suggestion that the symmetry criterion for comprehension might be taken to be related to the old idea that mathematical objects are constructed by abstraction from structures: the class of structures which are images under functions $j^n(f)$ of a class X might be taken to be an abstraction from the structure of X as an n-fold iterated collection.

The result does have some limitations.

It is not possible for impredicative class comprehension to hold in our class theory. If we required $[x \mid \phi]$ to exist for ϕ any formula in the internal language with quantifiers over C rather than S, then the definition of the class of Russell-Whitehead ordinals of true well-orderings (well-orderings satsifying the stronger condition that every *subclass* of the range rather than every subset of the range has a minimal element) would give the Burali-Forti paradox, because this class is unavoidably symmetric to the correct degree and equally clearly cannot be a set.

It is known that the class of strongly cantorian sets (see [**?**], pp. ??? for a discussion of this concept in NF-like theories) is both invariant under set permutations, and cannot consistently be a set. This allows us to deduce that there must be some non-set but setlike internal permutations in our class theory whose 2-action moves a strongly cantorian set to a non-strongly cantorian set, which is to say the least odd. It is easiest to see what happens if one considers the class of strongly cantorian Russell-Whitehead ordinals: one proves that for any set A there is a permutation $j^2(\pi)$ which fixes A and moves a strongly cantorian well-ordering to a non-cantorian well-ordering. This might be thought to give us some insight into what a model of this theory might look like.

It is important to note that we are not claiming to know how to construct

a model of predicative class theory (even with weak extensionality) satisfying the conditions of the Main Theorem.

6 Appendix: Symmetric Set Theory presented independently

We briefly present the axioms of the predicative theory of classes with the additional axioms we have proposed and required supporting definitions, producing a specification of what we will call "symmetric set theory", or, very briefly, SST. Here we allow ourselves to use $\in$ to represent the membership relation of the predicative theory of classes itself, since we are not mentioning the metatheory here.

The primitive relations of symmetric set theory are equality and membership ($x = y$ and $x \in y$). General objects of symmetric set theory are called classes.

axiom of extensionality: For all classes x, y,

$$x = y \leftrightarrow (\forall z : z \in x \leftrightarrow z \in y).$$

Alternatively, we could assume weak extensionality: for all classes x, y, z, if $z \in x$ and $(\forall u : u \in x \leftrightarrow u \in y)$, then $x = y$. In this case we specify a constant $\emptyset$ such that $(\forall x : x \notin \emptyset)$ and refer to elementless objects other than $\emptyset$ as atoms.

definition of sethood: We define $\mathtt{set}(x)$ as $(\exists y : x \in y)$. In English, we may say "x is a set". We define $(\forall x \in V : \phi)$ as $(\forall x : \mathtt{set}(x) \rightarrow \phi)$ and $(\exists x \in V : \phi)$ as $(\exists x : \mathtt{set}(x) \rightarrow \phi)$. We refer to these as bounded quantifiers, and we call a formula bounded iff all quantifiers appearing in it are bounded.

axiom scheme of predicative class comprehension: For any bounded formula ϕ (which may include parameters) and variable A not occurring in ϕ,

$$(\exists A : (\forall x : x \in A \leftrightarrow \mathtt{set}(x) \wedge \phi))$$

is an axiom. The witness to this axiom for a particular ϕ is unique by extensionality and may be written $\{x : \phi\}$: if weak extensionality is used and $\mathtt{set}(x) \wedge \phi$ is false for all x we define $\{x : \phi\}$ as $\emptyset$. We define V as $\{x : x = x\}$, and note that if we define $(\forall x \in A : \phi)$ as $(\forall x : x \in A \rightarrow \phi)$ and $(\exists x \in A : \phi)$ as $(\exists x : x \in A \wedge \phi)$ as is usual, the two apparently different definitions of quantifiers bounded by V agree.

axioms of set existence and pairing: There is a set. For any sets x, y, $\{z : z \in x \vee z \in y\}$ is a set, which we may write $\{x, y\}$. We will write $\{x\}$ instead of $\{x, x\}$.

definition of ordered pair: For any sets x, y, (x, y) is defined as $\{\{x\}, \{x, y\}\}$. It is straightforward to prove that $(x, y) = (z, w) \to x = z \wedge y = w$.

definition of relations, functions, permutations: A relation is a class of ordered pairs. A function is a class of ordered pairs f such that for any sets x, y, z, $(x, y) \in f \wedge (x, z) \in f \to y = z$. For any function f and set x, $f(x)$ is defined as the unique y (if there is one) such that $(x, y) \in f$. For any relation R, R^{-1} is defined as $\{(y, x) : (x, y) \in R\}$. A function f is an injection iff f^{-1} is a function. A function f is a permutation iff $f(x)$ is well-defined for each set x and f^{-1} is a function and is well-defined for each set x.

image and the j operation: For any function f and class A, $f\text{“}A$ is defined as $\{y : (\exists x \in A : y = f(x))\}$: if weak extensionality is used, the previous definition applies only if A has elements, and $f\text{“}A = A$ for every elementless object A. For any permutation f, we define $j[f]$ as $\{(x, f\text{“}x) : x = x\}$. Observe that $j[f](x) = f\text{“}x$ iff $f\text{“}x$ is a set. Observe further that $j[f]$ will be injective if it is total (defined at all sets), and a permutation if $j[f^{-1}]$ is total as well. For each concrete natural number n, we define $j^n[f]$: $j^0[f] = f$ and $j^{n+1}[f]$ is defined as $j[j^n[f]]$ if $j^n[f]$ is a permutation. A permutation f is said to be n-setlike iff $j^n[f]$ is a permutation. Note that the n in this notation is not a variable that we can quantify over.

symmetry defined: We say that a set x is n-symmetric iff there is a set s (called a support for x) such that any $(n-1)$-setlike permutation f such that $j^{n-1}[f](s) = s$ we also have $j^{n-1}[f]\text{“}x = x$.

axiom of symmetric set comprehension: A class is a set iff it is 3-symmetric: i.e, a class x is a set iff there is a set s such that for any 2-setlike permutation π such that $j^2[\pi](s) = s$ we have $j^2[\pi]\text{“}x = x$.

The results of the earlier part of the paper show us that each axiom of NF [or NFU if weak extensionality is used] can be proved in this theory, if each quantifier is taken as restricted to V and each parameter is taken to refer to a set.

7 Appendix: sketch of a partial converse result

This work is based, as is noted above, on an adaptation of a theorem of Forster and others, to the effect that the stratified sentences are exactly the sentences which are invariant under setlike permutations in all Rieger-Bernays models. This theorem was actually the motivation for all the work in this paper: it gave a semantic motivation for the notion of stratification, which I set out to adapt into a semantic motivation for the axiom of stratified comprehension.

The class of sentences we consider is a little more general. We allow a stock of function symbols representing setlike permutations. The rule for these in relation to stratification is that any occurrence of a term must have the same setlike permutation (or none) applied to it in all contexts in which it appears (where terms are either variables or a function symbol applied to a term). Terms are then assigned relative type in the usual way, and any term must appear with the same relative type wherever it appears in a membership or equality formula. We also provide constants, which do not need to be assigned type.

The notion of invariance that we employ is then that a formula $\phi(x_1, \ldots, x_n, a_1, \ldots a_n)$ is invariant iff there are numbers τ_i, σ_j such that for any allowable permutation g such that $j^{\sigma_1}[g](a_1) = a_1, \ldots, j^{\sigma_n}[g](a_n) = a_n$ we have $\phi(x_1, \ldots, x_n, a_1, \ldots a_n) \leftrightarrow \phi(j^{\tau_1}[g]x_1, \ldots, j^{\tau_n}[g]x_n, a_1, \ldots a_n)$. We claim that any such formula is equivalent to a stratified formula (in the general sense indicated above).

We have already shown above why any stratified formula is invariant in this sense: allowing variables to be adorned with additional function symbols in the way indicated will not change this.

Let (M, ϵ_M) and (N, ϵ_N) be two structures for the language of set theory, satisfying [weak] extensionality. Suppose that they satisfy the same stratified sentences. Transform them into models of type theory (M_i, ϵ_M^τ), (N_i, ϵ_N^τ) for the language of type theory, in which each type M_i is just a copy $M \times \{i\}$ ofr M and $(x, i)\epsilon_M^\tau(y, i+1)$ iff $x\epsilon_M y$. The models of type theory will of course still satisfy the same stratified sentences.

Use a back and forth argument to construct models $(M_i^*, \epsilon_{M^*}^\tau)$, $(N_i^*, \epsilon_{N^*}^\tau)$ which are elementary extensions of (M_i, ϵ_M^τ), (N_i, ϵ_N^τ) respectively and which are isomorphic as models of type theory (ignoring the relation of having the same first component in M or N). Where we have not yet created the image

under the isomorphism f of $(x,i) \in M_i$ (or later of $x \in M_i^*$) we first give a complete description of the theory of (x,i) (including a complete description of its first component in terms of the original theory of M). We then create a new object (or use an existing one if it is determined already) having the correct theory to be $(f(x),i)$ (this is possible because the two structures have the same stratified theory) then add the new object to the model N^* (with a complete description of $f(x)$ as an element of N^*, including its first component in N), adding analogues $(f(x),j)$ in every type. Then choose the next element of the N or N^* model and define an image under f^{-1}. Continue this process back and forth until it terminates in a pair of models with an external isomorphism. This isomorphism will in terms of M^* and N^* give a setlike permutation from a model elementarily extending M to a model elementarily extending N. This establishes that models with the same stratified theory will in fact satisfy the same invariant sentences and vice versa. For any model M satisfying a particular invariant sentence ϕ [suitably decorated], the conjunction of all stratified sentences which must hold in any model satisfying ϕ is logically equivalent to ϕ by completeness and must be equivalent to a finite subconjunction by compactness.

Now to build a model of SST [or SSTU] from a model of NF [or NFU], subject to the additional condition that each set in the model of NF(U) is 3-symmetric relative to set permutations, proceed by listing and processing the definable classes paired with candidate supports: if a definable class is not a set, and so not defined by a stratified formula, there will be a model (constructed as above) in which there is a setlike permutation moving it and fixing the candidate support. Pass to this model (an elementary extension of the previous model) and add the setlike permutation to the language, adding new formulas and support elements to process later. Notice that since the new model is an elementary extension of the previous model, previously established supports of sets will be preserved at each step. Continue. This process will terminate (even if it leads to a class model, but I believe it will actually terminate in a set model) in a model which satisfies SST. It will have the same NF [or NFU] theory as the original model, but the embedded NF model may be much larger than the original one (and have lots more external subclasses than the original one).

NOTE: I do not fully believe this. I believe that every definable class which is not a set (and so cannot be stratified, and so cannot be invariant) can be arranged not to be symmetric, but I do not see how it enforces symmetry of sets in the model.

The problem is that we want to have at least that every set in the model of NF is 3-symmetric with respect to all setlike permutations of the model (that has to be true for us to have any chance). But can we add new setlike permutations while preserving this condition? We know exactly how we add new permutations: the image of an object under a new permutation. New idea: add a primitive predicate "x is a support of y": in the initial model, this means that x witnesses that y is 3-symmetric with respect to setlike permutations of the original model. Then when constructing isomorphisms between models of the same stratified theory, follow the rule that each new permutation introduced must respect supports in the obvious sense; I need to check that the back-and-forth construction of permutations actually goes through with this additional condition, but it seems reasonable that it should. I'm still not convinced that this works, because I'm worried that new definable setlike permutations might be introduced as the model is augmented.

This doesn't give a conservative extension result: the additional condition is required, as noted above, that each set in the original model of NF(U) was symmetric. We do not claim here to know how to construct a model of NF, and even if we did, it is unclear how to construct a model of NF in which every set is symmetric in the relevant sense. We do know how to construct models of NFU but we do not know how to construct a model of NFU in which each set is symmetric in the relevant sense.

We repeat our comment that this theory appears to have a motivation in rather old-fashioned terms: the idea is that mathematical objects are understood as being constructed by **abstraction**. Passing from a class to its symmetrization relative to setlike permutations fixing a previously given set looks like abstracting a feature from the structure of the class.

8 Appendix: random notes

Question: is it a theorem of NF, or perhaps of NFU + Choice, that all sets are 3-symmetric with support? Such a result seems unlikely, but it would give a conservative extension result by the argument above.

Note: it is interesting to investigate how choice fails in this theory on its own terms. There clearly cannot be a true well ordering of the universe: if there were such an order, the set of segments in the order would be rigid (any permutation fixing it would fix everything) and so would serve as a support for the Russell class, leading to absurdity. The nonexistence of a set well-

ordering is harder to prove: of course it is provable using the Main Theorem and the Specker disproof of AC in NF, but it would be nice to see how to prove it directly from symmetry. It is not clear that a linear order on the universe is impossible in SST.

blue sky idea: Suppose X is a set for which we are trying to construct a support. Consider the orbits under the set of permutations $j^2(\pi)$ such that $j^3(\pi)$ fixes X of elements of X (and possibly also of non-elements of X). Suppose that there are no more than $T^2(|V|)$ such orbits, so that each one can be correlated with a double singleton. Let S_1 be the set of all singletons belonging to double singletons correlated with orbits included in X. Let S_2 be the set of all Quine pairs of a singleton with an element of an orbit correlated with the singleton of that singleton. If a permutation $j^2(\pi)$ fixes both S_1 and S_2, then $j^2(\pi)$ sends elements of orbits in X to elements of orbits in X, so $j^3(\pi)$ fixes X. The cardinality hypothesis which makes this work is pretty strong (I have no idea whether this is actually possible!). The argument as presented here seems to require strong extensionality, though it might be adaptable to the NFU case.