A. CARBONE and S. SEMMES

# LOOKING FROM THE INSIDE AND FROM THE OUTSIDE

ABSTRACT. Many times in mathematics there is a natural dichotomy between describing some object from the inside and from the outside. Imagine algebraic varieties for instance; they can be described from the outside as solution sets of polynomial equations, but one can also try to understand how it is for actual points to move around inside them, perhaps to parameterize them in some way. The concept of formal proofs has the interesting feature that it provides opportunities for both perspectives. The inner perspective has been largely overlooked, but in fact lengths of proofs lead to new ways to measure information content of mathematical objects. The disparity between minimal lengths of proofs with and without "lemmas" provides an indication of internal symmetry of mathematical objects and their descriptions. A principal observation of this paper is that mathematical structures can be embedded into spaces of logical formulae and inherit additional structure from proofs. We shall look at finitely-generated groups, rational numbers and $SL(2, \mathbf{Z})$, and examples from topology and analysis.

Why can a logical formula be "hard to prove"? *Truth* and *provability* are both connected to questions of algorithmic complexity. The existence of a polynomial time algorithm for deciding whether a propositional formula is a tautology or not is equivalent to the famous $P = NP$ problem. The existence of polynomial-size proofs for all propositional tautologies is related to the complexity problem *NP* = *co-NP*. In predicate logic the set of tautologies is not algorithmically decidable among all formulae, and this implies that minimal proofs can have non-recursive size as compared to the size of the tautology.

It is not only the *absolute* size of proofs which matters. The minimal size of a proof of a given formula may be much shorter when one is permitted to use "lemmas" than when one is required to give a "direct" proof. These notions are made precise in formal logic through the rules of *modus ponens* and *cut*. Concrete examples with explicit lower bounds are given in (Orevkov 1982; Orevkov 1993; Statman 1974; Statman 1978; Statman 1979; Tseitin 1968). The exposition (Carbone and Semmes 1997) provides an introduction to these matters.

The existence of indirect proofs of a given formula which are much shorter than the direct proofs seems to reflect the presence of some kind of *symmetry* in the underlying mathematical constructions and descriptions.

*Synthese* **125:** 385–416, 2000.

This principle is not at all understood. It is not even clear how to formulate it precisely. In this paper we explore these issues and try to bring some of the ideas of formal logic closer to ordinary mathematics. The *lengths* of proofs provides a way to measure the "information" content in mathematical objects. If one used only *direct* proofs this would, roughly speaking, be just a measurement of the size of an explicit construction. By allowing *lemmas* the lengths of proofs yields a more subtle measurement of information which has not been studied as such. In fact, one can make a variety of different measurements by imposing restrictions on the logical nature of the lemmas.

In an indirect proof one can code many substitutions implicitly which would have to be carried out explicitly in an actual construction. Restrictions on the logical nature of the lemmas – such as the number of nested quantifiers – lead to restrictions on the complexity of the substitutions. The efficiency of the coding of the substitutions is a reflection of *internal symmetry*.

These issues come out clearly in the context of *feasible numbers* as studied in (Carbone 2000a). There one sees *cycling* in proofs and the way that this cycling is related to the logical nature of the lemmas. In this context of natural numbers one can have a lot of cycling in short proofs of feasibility, and a great deal of compression in these proofs. The length of the shortest proof of the feasibility of a given number provides a measurement of information, as above. More generally, one can use the *idea* of feasibility to code mathematical constructions into formal proofs and therefore as a tool for measuring information content in new ways. Note that implicit descriptions of constructions provided by proofs with lemmas can be converted into explicit constructions through the procedure of "cut elimination" (Girard 1987b; Takeuti 1975; Carbone and Semmes 1997), but this can lead to large expansion. Related matters concerning feasibility, formal proofs, and cut elimination are discussed in (Carbone and Semmes 1999; Carbone and Semmes 2000).

The general idea of feasibility can be applied to *finitely-generated groups* (as in Section 6 below). Again we think of the length of the shortest proof of the feasibility of an element of the group as a measurement of its "information" content. This case is very different from that of the natural numbers because of the way that group relations can affect the "information" in the elements of the group. The usual word metric is also sensitive to the group relations, but using the lengths of proofs one can take into account additional symmetry. For instance, in cyclic subgroups the measurements coming from lengths of proofs are already nontrivial, independently of the relations. One does not get as much compression as for the natural numbers (which have more "symmetry"), but some of the arguments for natural numbers carry over to groups. Thus even for free groups the word metric is different from "lengths of proofs". They are different in the treatment of asymptotic directions, and in the distinction between those that follow cyclic subgroups and those that do not.

This approach to groups was motivated in part by (Carbone 2000c). There one starts with a proof, considers its *logical flow graph*, which traces the flow of occurrences of formulae in a proof (see (Buss 1987; Carbone 1997)), and then associates to the graph a finitely-presented group by reading the cycles in the graph. Short proofs of the feasibility of large numbers, for instance, lead to strongly distorted groups. In this paper we go in the reverse direction, from groups to proofs.

We shall consider the idea of feasibility in the context of *rational numbers*. One can imagine that these measurements of information can be related to number-theoretic properties, but we do not know any concrete results at this time. Short proofs of feasibility of complicated numbers can be facilitated through natural dynamical systems, such as the action of $SL(2, \mathbf{Z})$ on $\mathbf{Q}$ by projective transformations. That is, the methods applicable to abstract groups can also be used for groups of transformations on $\mathbf{Q}$.

We shall consider well-known constructions in *topology*. In spirit these constructions are compatible with the idea of feasibility. We shall describe an example of a *torus bundle* in which there is exponential distortion induced by a simple cycling which is similar to the cycling and substitutions that can occur within proofs. We shall also look at ordinary differential equations and exponentiation with continuous parameters.

Topological notions do not lend themselves to such clear formalization as for groups and numbers, but they also serve other purposes. We mentioned before that basic questions of complexity theory have equivalent formulations in the context of logic, and also that there can be a large gap between the minimal size of proofs without lemmas of a given statement and proofs with lemmas. There is no simple model to explain when the minimal size of a proof or the relative sizes of proofs with and without lemmas should be polynomial or exponential. In the topological setting there are examples with intricate structure but simple geometric explanation which might be indicative of more general models. In particular the exponential growth which can occur has a nicely geometric realization.

Ordinary mathematics should be seen as a potentially rich resource of interesting examples for the study of formal proofs, and one of the purposes of this paper is to bring out some of these examples, and the general way in which the concept of feasibility can be used to this end.

One should compare the present discussion of information content with Kolmogorov complexity and algorithmic information theory. (See (Kolmogorov 1968; Chaitin 1987; Chaitin 1992; Li and Vitányi 1990), for instance.) Part of the point here is to loosen the degree of implicitness, and consider intermediate measures suggested by logical derivations.

Before we proceed to more precise discussions of information and complexity, let us begin with some basic questions concerning the nature of descriptions.

## 1. INNER AND OUTER DESCRIPTIONS

What is a set? How can a set be described? These are basic questions which reverberate in mathematics down to the foundations. Let us consider them here in the practical way of what mathematicians actually do.

There is a basic distinction between what one might call *inner* and *outer* descriptions of sets. For an outer description one might have a given set $A$ embedded into some larger space $X$ of simple structure, and one may describe $A$ by specifying rules which determine which elements of $X$ are in $A$. An inner description might provide a listing of the elements of $A$, with more concern for the internal structure of $A$ than an embedding of it into a larger space.

Let us consider an example. How can we describe a curve $\Gamma$ in the plane? One answer might be to provide a parameterization of it, $(x(t), y(t))$, $t \in \mathbf{R}$. Another possibility is to define $\Gamma$ as the set of solutions to some equation,

$$\Gamma = \{(x, y) \in \mathbf{R}^2 : F(x, y) = 0\},$$

where $F(x, y)$ is some function.

These are very different ways to describe a curve. In the first case it might be easy to generate many points on the curve without having a general understanding or test for when a point lies on it. For inner descriptions it may not be clear how many points are needed to have a reasonably accurate picture of the set in question, and one may have to be careful about exploring well one part while missing another. In the second case one might have a simple characterization of the elements of the set without a clear idea of how to find actual solutions.

Consider the case where we define $\Gamma$ as the zero set of $F(x, y)$ with $F$ a polynomial. A basic point about the algebraic notion of a plane curve is that it may not be compatible with the notion of a parameterization. Over the real numbers the zero set might be empty, or have several components, including compact components, etc. Some of these problems can be alleviated by working with complex numbers and making assumptions of irreducibility. A more interesting incompatibility with the idea of a parameterization is that the curve might not be rational, so that it may not be reasonable to try to parameterize the curve with the ground field. It might be an elliptic curve or a curve of higher genus. (See (Kirwan 1992) for a discussion of algebraic curves.)

If we work over a field like the rational numbers, there might be more basic problems about the existence of points on the curve. (Someone once pointed out to us that a great idea in algebraic geometry is that one can study sets of equations independently of whether one knows that there are solutions. Part of the point is that the underlying choice of field can change, and the sets of solutions with it. The set of solutions might be empty for one field, and then be quite large for another field.)

Instead of thinking algebraically, we can think more in terms of calculus. We should be careful about what kind of functions we allow, though. For instance, any closed subset of the plane can be realized as the zero set of a $C^\infty$ function $F$. (Take $F(x, y) = \exp(-\Delta(x, y)^{-1})$, with $\Delta(x, y)$ a function on the plane which gives a regularized version of the distance to the closed set, as on page 171 of (Stein 1970).) Thus the $C^\infty$ property is too flexible by itself to provide a practical way to describe sets. We can avoid this problem by restricting ourselves to smooth functions $F$ whose gradient $\nabla F$ does not vanish on the zero set of $F$. This is the hypothesis of the implicit function theorem, which then implies that the zero set of $F$ is locally given by a smooth curve. One can have singularities at critical points, as in the case of polynomials, and there are theories for analyzing these.

This is a very basic example which hopefully illustrates well what we have in mind by “inner” and “outer” descriptions. We also see how the context matters. It is very different to think algebraically in terms of polynomials than in terms of more general functions. Calculus permits a more flexible idea of function, while algebra is more rigid in some ways, but enjoys more flexibility of context, in that one might switch to the rationals or other ground fields.

If we want to make an inner description by listing points, one can ask that this listing respect the structure of the situation, like smoothness or algebraic properties. In topology one would normally impose continuity conditions on mappings, and so on.

These ideas show up in many different contexts in mathematics. For instance, one can take some finite set $\mathcal{A}$ as an alphabet, and look at the set $\mathcal{A}^*$ of all *words* generated by elements of $\mathcal{A}$, i.e., all finite strings of elements of $\mathcal{A}$. One might have a *language* $L$ based on $\mathcal{A}$, which is to say a subset of $\mathcal{A}^*$. A priori $L$ could be anything. How might it be described? $L$ might be *effectively enumerable*, so that there is an algorithm for generating all of its elements. This is a kind of inner description. Instead there might be an outer description, like an algorithm that says when a word lies in $L$.

Roughly speaking, inner descriptions correspond to ways to produce effective witnesses, while outer descriptions correspond to ways to check membership and to decide yes/no questions.

If one knows that $L$ is pretty "thick" – i.e., one has reasonably large lower bounds on the number of elements of $L$ among the words of length $n$ – then one might be able to get a practical way to list the elements of $L$ from the algorithm for deciding whether a word lies in $L$ or not. One simply goes through all the words and keeps the elements of $L$. This need not work very well if $L$ is too sparse. In the case of integer solutions of a polynomial equation $F(x, y) = 0$, it may be very difficult to tell if there are any solutions (Hilbert's tenth problem) or to know how many. In general the existence of integer solutions of polynomial equations is algorithmically undecidable, but this is not known for the rationals or other fields.

Conversely, there are sets which are effectively enumerable but for which there is no algorithm to decide membership.

There are nice variations on this theme of thickness and sparseness of languages in the context of the P = NP problem. See (Johnson 1990, 87).

As another example, suppose that we have an $n \times n$ matrix of complex numbers, which we think of as defining a linear mapping $T$ on $\mathbf{C}^n$. A complex number $\lambda$ is an *eigenvalue* of $T$ if $\lambda I - T$ is not invertible as a linear mapping on $\mathbf{C}^n$. The set of eigenvalues is called the *spectrum* of $T$. One can define it more concretely as the set of zeros of the polynomial equation

$$\det(\lambda I - T) = 0.$$

This is a perfectly good definition of the spectrum, but how does one actually find eigenvalues? This is a tricky question whose numerical solution is of great importance and much studied.

Dynamical systems provide another interesting case to consider. One might be able to generate a good approximation to an attractor quickly from the inside, looking at iterates of a critical point for instance, while the "rules" which govern the geometry of the attractor might be hard to see. The number of points needed to have an accurate picture of the attractor might be unclear as well.

Inner and outer descriptions need not be very compatible with each other. In mathematics one is often much more accessible than the other. Which one is more accessible can depend on the context.

In the next section we shall discuss the particular case of the set of all tautologies inside the space of formulae. In Section 3 we discuss how individual formulae can in turn be used to describe subsets of other sets (equipped with some structure), and in Section 4 we consider the general relationship between algebraic structures and points inside a set.

In Section 5 we take up the notion of *feasibility*. This provides a way to embed a mathematical structure inside the space of formulae (with respect to some language). The combinatorics of formal proofs then induces *new* structure on the original mathematical object. For this the *cut rule* is particularly relevant. This idea is developed through examples in Sections 5–10.

This paper is intended to be accessible to a broad audience. Readers not very familiar with formal proofs may find (Carbone and Semmes 1997) a useful source of background material. See also (Carbone and Semmes 1999; Carbone and Semmes 2000) for some connected topics.

## 2. THE SET OF TAUTOLOGIES

The set of tautologies provides an interesting case to consider for inner and outer descriptions. One can consider either propositional or predicate logic.

Imagine fixing a collection of variables and the rest of a logical language, so that one has specified a notion of formulae. Let us think of the set of all formulae as being relatively simple (e.g., a recursive set), and imagine that we are interested in understanding the set of all tautologies as a subset of it through both inner and outer descriptions.

For an inner description of the space of all tautologies we can use *proofs*. The rules for building proofs provide a way to move around in the space. It may not be easy to reach a particular tautology, but in principle we can go anywhere in the space through proofs.

Given two tautologies we can make a new one through binary logical rules. There are numerous ways in which to wander around in the space as a whole. The structure of the possible ways to move within the space reflects its geometry.

The idea of the relationship between the geometry of a space and the ability to move around in it is much studied in other parts of mathematics.

What about outer descriptions? We can use *semantics* to provide a kind of outer description of tautologies. The completeness theorem says that the set of provable formulae is the same as the set of formulae which are “true” in all interpretations. We can think of each interpretation as a test. Although there are many such tests (and indeed the set of predicate tautologies is algorithmically undecidable, under modest conditions on the language), it is remarkable nonetheless that tautologies enjoy these outer and inner descriptions simultaneously. One can argue that it is reasonable that neither description is very simple given that we are lucky enough to have both.

In the case of propositional logic some of these issues emerge more clearly. The "outer" characterization of tautologies as being the formulae which are true in every interpretation implies that the set of tautologies is co-NP. If P = NP (and hence P = co-NP) then there is a polynomial-time algorithm which tells whether a propositional formula is a tautology. This would be a very effective outer description.

It is not known exactly how the size of a tautology is related to the size of its shortest proof. The existence of short proofs is a way to say that the *inner* description of propositional tautologies through proofs is efficient.

See (Urquhart 1987) concerning some interesting structure connected to "hard" examples of propositional tautologies and interpretations.

Propositional and predicate logic provide very basic examples of sets in mathematics whose descriptions one would like to understand better. Another interesting example is provided by Brouwer's intuitionistic logic. In this system disjunctions and existential quantifiers are treated differently from classical logic. One cannot assert $A \vee B$ without actually having a proof of one of $A$ or $B$; in particular, one does not take $A \vee \neg A$ as being automatic. Similarly, one cannot assert $\exists x R(x)$ without having a proof of $R(t)$ for some concrete term $t$. The structure of proofs is somewhat simpler in this case than in classical logic, but the notion of interpretations for characterizing tautologies (in intuitionistic logic) is more complicated. (Tautologies in intuitionistic logic are always tautologies in classical logic, but the reverse is not true.)

We should mention that this is the same Brouwer who proved the famous fixed-point theorem.

Logical formulae can describe or involve mathematical objects. The existence of short proofs leads to efficient inner description for the set of tautologies. A proof of a formula can reflect the structure of the underlying objects. Intuitionistic and classical logic differ both in the way that they describe mathematical objects, and in the way that their sets of tautologies are described. For Brouwer a proof is a kind of function, where a rule like modus ponens corresponds to composition of functions. This is connected to the theory of Lambda Calculus, which associates functions to intuitionistic proofs in a way that reflects their internal structure.

## 3. DESCRIBING SETS THROUGH LOGICAL FORMULAE

Mathematical logic provides interesting ways to make descriptions of sets. The most basic method comes from model theory for first-order logic. The reader who is not familiar with these concepts need not lose heart, we simply want to have an impression in mind.

With this method, one can talk about structures and defining special subsets of a given set abstractly, independently of any specific set. Before we say what this means in general, let us think about groups. There is an abstract idea of groups that exists independently of any particular group. There is also a way to talk about certain subsets of a group, like the subset of elements of a certain order, or the center of the group, that exists independently of any particular group.

The idea of abstract mathematical structures can be formalized through a logical language and a set of axioms for the objects involved. In a first-order language one has the usual logical connectives which represent "and", "or", "not", "implies", and the quantifiers "for all" and "there exists", and there are additional symbols which reflect the particular structure. These are symbols for variables, constants, functions, and predicates (relations). Each function symbol and predicate has a fixed number of arguments, called the arity. The number and the arities of the function symbols and relations depend on the given mathematical structure.

For example, for the theory of groups one uses one relation, the binary relation of equality =. One can use two function symbols, one binary and one unary, which correspond to group multiplication and inversion. A separate function symbol for inversion is not really needed, though, because existence and uniqueness for inverses can be given through the group axioms. There is one constant symbol, corresponding to the identity element of the group.

These are all just symbols, however, with no underlying set. This is because a first-order language concerns the *idea* of a group rather than a particular one.

One also needs the notion of a *term*, which is an expression constructed from variables and constants using function symbols. Think of a formal expression for groups, some product of variables, possibly with inverses. The functions and relations take terms for their arguments, as in the composition $s^{-1}t$ and the relation $s = t$ in the context of groups, where $s$ and $t$ are terms.

A relation with a choice of arguments – like $s = t$ – is a logical formula in a first-order language, an atomic formula. Informally, it is a statement which might be true or false, depending on the context (such as a particular group). These atomic formulae can be combined with the logical connectives to build more complicated formulae.

The *theory* of groups is given by the usual axioms governing the group operations, such as the associativity axiom. One can have other theories, based on different languages or axioms.

All of this exists purely at the level of formal symbols. Roughly speaking, a *model* is a specific choice of a set and interpretations in or on it for the constants, functions, and relations. The variables would be interpreted as taking values in the set. The choices for the constants, functions, and relations should satisfy the axioms of the theory. Thus actual groups are models for the first-order theory of groups, with a set of group elements, the usual notion of equality, a choice of group operation, etc. There are many different kinds of groups, many different models, but one first-order theory of groups.

Another example is provided by arithmetic. One can formalize it with the binary relations $=$ and $<$, operations like addition and multiplication, and the well-known Peano axioms. The usual notion of natural numbers provides a model for this theory, but there are nonstandard models too.

A first-order language provides the possibility to make universal recipes for describing certain sets, a set for each choice of model. Each formula in the language defines such a recipe. If $\phi(x, y, z)$ is a formula, with free variables $x$, $y$, and $z$ and no others (for instance), and if we have a specific model based on a set $S$, then we get a subset of $S \times S \times S$, namely the set of triples $(x, y, z)$ for which $\phi(x, y, z)$ is a valid formula. For example, one can define the center of the group in this way, or the set of elements of order 2.

One can also define sets that depend on the particular model, by using specific elements of the underlying set $S$ in the definition. This can be viewed in terms of the situation in the preceding paragraph, by taking a *section* of the type of set defined above. That is, one can assign particular values to some of the free variables, to get a subset of a Cartesian product of $S$'s, with the number of factors in the Cartesian product reduced from what it was before by the number of free variables to which values are assigned.

One can think of these as ways to make "outer" descriptions of certain classes of sets in terms of logic. It is rather sophisticated, because of the possibility of quantifiers. Without quantifiers it is already tricky, but quantifiers make it even more complicated. Many problems of algorithmic decidability of sets involve finding a uniform way to eliminate quantifiers. In the description of some sets, quantifiers cannot be eliminated, and even when quantifiers are eliminable, the length of the quantifier-free description often becomes extremely large, and difficult to handle.

## 4. SOME COMMENTS ABOUT ALGEBRA AND POINTS

A common phenomenon in algebra is to have algebraic structures which make sense abstractly, but which arise classically in more geometric ways, involving points in sets. A fundamental example is given by groups, with the abstract notion of groups on the one hand, and groups of transformations on sets on the other. Another basic example is the following. Let $X$ be a compact Hausdorff topological space, and let $C(X)$ denote the space of all complex-valued continuous functions on $X$. This is an algebra, which is commutative, and even a $C^*$-algebra. One can also talk about algebras abstractly, independent of some kind of realization on a space like this.

If $X$ and $C(X)$ are as above, then one can recover points in $X$ from the algebraic structure in $C(X)$. If $p$ is a point in $X$, then $\{f \in C(X) : f(p) = 0\}$ is a maximal ideal in $C(X)$, and conversely, every maximal ideal arises in this manner. Given another compact Hausdorff space $Y$, one can use this fact to show that $X$ is homeomorphic to $Y$ if and only if $C(X)$ is isomorphic to $C(Y)$ as an algebra. One can characterize the algebras that arise this way, as commutative $C^*$-algebras. See (Rudin 1991; Simmons 1963). There are analogous stories in the context of algebraic varieties, but let us stick to topological spaces for simplicity.

In principle, compact Hausdorff spaces are described completely by the algebra of commutative functions on them, but how does this work practically? How can one see inside the space through the algebra? In some kind of practical way, and not just in principle? This turns out to be subtle and mysterious. There is a different way to try to represent the structure of a space in purely algebraic terms, through which one can recover topological invariants of the underlying space from direct algebraic constructions. See (Connes 1994). This approach also gives meaning to these topological invariants in non-commutative settings where there need not be "points" in the classical sense, and this is a matter of great current interest (in a number of directions, including mathematical physics).

This is similar in spirit to the relationship between operational and denotational semantics in programming languages. See (van Leeuwen, 1990), for instance.

It can happen naturally that one has an algebra in hand, but not the underlying points that one might want (at least not directly). For example, let $T$ be a linear transformation acting on some $\mathbf{C}^n$. Consider the algebra of linear transformations generated by $T$, which amounts to saying all (complex) polynomials in $T$. For this we include the identity transformation on $\mathbf{C}^n$, which one can think of as $T$ to the 0th power, and which is associated to the constant polynomial equal to 1. This is a nice commutative algebra, but what are the underlying "points"?

Suppose that $T$ is defined by a diagonal matrix, with diagonal entries $\lambda_1, \ldots, \lambda_n \in \mathbf{C}$. These diagonal entries are then the "points" in a natural way. If we let $X$ denote the set of them, then there is a simple correspondence between the algebra of linear transformations generated by $T$ and the algebra of restrictions of complex polynomials on $\mathbf{C}$ to $X$. The latter is isomorphic to the algebra of complex functions on $X$. Note that some elements of $X$ may occur more than once as a diagonal entry of $T$.

If $T$ is not diagonal, but is diagonalizable, then there is a similar correspondence for the algebra that it generates, even if this might not be obvious at first glance. The diagonal entries in the diagonalized form are the same as the eigenvalues of $T$. If $T$ is not diagonalizable, but has a Jordan canonical form with nonzero parts off of the diagonal, then the notion of "points" underlying the algebra is more complicated, because of nilpotency. In particular, the restrictions of polynomials on $\mathbf{C}$ to the set of eigenvalues does not tell the whole story, and one does not reduce to functions on this set in the end. The nilpotency leads to extra structure around some (and maybe all) points in the spectrum.

There are versions of this for linear operators acting on infinite-dimensional spaces, in which the natural notion of spectrum is a set which may be infinite, and whose topological structure becomes important. See (Rudin 1991). This is connected to the earlier discussion, concerning the algebra of continuous complex-valued functions on a topological space. There are also issues of diagonalizable versus non-diagonalizable linear operators, so that there can be more structure involved than just the spectrum as a set of points, or as a topological space, as in the case of finite matrices.

Another basic situation concerning algebraic structures and underlying "points" is provided by Boolean algebras. In this case, the connection with points is somewhat simpler, as in Stone's theorem. (See (Halmos 1974).)

In proofs, there is a kind of algebraic structure involved, and it is also natural to think of proofs in terms of sets with points (such as atomic formulae) and combinatorial structure. This is a remarkable coexistence.

This idea is illustrated by (Carbone 2000b), in which the *Craig interpolation theorem* (Craig 1957) is discussed in a combinatorial context without the algebraic structure of connectives. This combinatorial view of points in a proof is also present in the notion of *logical flow graphs* (Buss 1991; Carbone 1997), which trace the logical connections within a proof, and in the study of *cycles* in these graphs, as in (Carbone 1997; Carbone 2000a; Carbone 2000c). Logical flow graphs are related to the earlier notion of *proof nets* (Girard 1987a). The concept of *inner proofs* from (Carbone 1997) gives another reflection of the idea of points moving around inside proofs.

## 5. FEASIBLE NUMBERS

There has been much concern in mathematics about abstraction which may not reflect anything concrete or “real”. Extremely large numbers were troubling to some, and there was the idea that they should be treated differently from a small number like 37 which is closer to ordinary existence.

The first mathematical treatment of *feasible numbers* was given in (Parikh 1971). (The philosophical discussions go back to Mannoury, Poincaré, and Wittgenstein.) For this we start with the first-order theory of arithmetic, and we add a unary predicate $F$. Roughly speaking, $F(x)$ is interpreted as meaning that $x$ can be constructed in some feasible manner. We shall use the arithmetic operations $+$ (addition), $*$ (multiplication), and $s$ (successor). In addition to the usual axioms of arithmetic, we add the following axioms for $F$:

$$\begin{array}{ll} & F(0) \\ F: equality & x = y \to (F(x) \to F(y)) \\ F: successor & F(x) \to F(s(x)) \\ F: plus & F(x) \wedge F(y) \to F(x+y) \\ F: times & F(x) \wedge F(y) \to F(x * y) \end{array}$$

In other words, 0 is considered to be feasible, and the property of feasibility is closed under equality, successor, addition, and multiplication.

For this discussion we do not permit ourselves to use induction over $F$-formulae. Otherwise we could prove $\forall x F(x)$ in a few steps. Note that if we add the axiom $\exists x \neg F(x)$, asserting the existence of a nonfeasible number, then we still get a consistent system, for which the models are nonstandard models of arithmetic.

The idea instead is that if we can write down a proof of $F(t)$ for some term $t$, then that should mean that $t$ was “feasible” in a reasonable sense. Of course we can always prove $F(n)$ for any natural number $n$ in about $n$ steps, using the successor rule repeatedly. (Strictly speaking, we are abusing the first-order language of arithmetic here, and $n$ really means the result of applying $n$ times the successor function to 0. Syntactic technicalities can detract from the main points, and we shall generally not pay attention to them here.) However, we can use the size of a proof of $F(n)$ as a measurement of the feasibility of $n$.

This is an appealing point. We can use proofs to make descriptions of mathematical objects, and to make measurements of their complexity. We shall leave aside the foundational issues and simply use the idea of feasibility as a tool for studying mathematical structures.

To make precise the measurements one should be careful about the formalization of proofs. We shall not discuss this in detail, but there are a couple of important points. The first is that we consider only proofs in which the result of any intermediate step is used only once. Thus proofs have tree-like structures. The second concerns the role of the "cut" and contraction rules in sequent calculus and their counterparts in other systems. Roughly speaking, the cut rule allows indirect reasoning through lemmas. It is a generalization of the deduction rule modus ponens, which says that if you know $A$ and if you know that $A$ implies $B$, then you can conclude $B$. Without the cut rule, a proof of $F(t)$ for some term $t$ would have to exhibit an explicit construction of the term $t$. The contraction rule allows multiple occurrences of a formula to be combined into a single one. In combination with cuts, contractions permit a piece of information to be used several times. With cuts and contractions, one can make short proofs of feasibility which provide only implicit descriptions, as we shall soon see. There are effective methods for converting proofs with cuts into proofs without them, at the cost of (possible) great expansion in the proofs. See (Girard 1987b; Takeuti 1975; Carbone and Semmes 1997).

Let us mention one more point. In (Parikh 1971), an $F$ : *inequality* axiom is included, to the effect that if $y$ is feasible and $x < y$ then $x$ is also feasible. For the historical concern about large numbers this is a reasonable requirement to consider, but we have omitted it intentionally. It does not fit as well with the idea of a proof of feasibility of $F(t)$ as providing a description of $t$, and it is less convenient for other mathematical contexts.

Let us now consider the concrete matter of how one might give short proofs of feasibility of numbers. We follow the examples in (Carbone 2000a).

As above, one can get a proof of $F(n)$ in about $n$ lines through repeated use of the $F$ : $successor$ axiom. Using the $F$ : *times* axiom repeatedly instead, one can get a proof of $F(2^n)$ in about $n$ lines.

Here is another method. We know that

$$(1) \qquad F(x) \to F(x^2)$$

by the $F$ : *times* axiom. In particular,

$$(2) \qquad F(2^{2^j}) \to F(2^{2^{j+1}})$$

for all $j = 0, 1, 2, \ldots$ We can combine $n - 1$ copies of (2) together with the feasibility of 2 (i.e., $F(s(s(0))))$ to get a proof of $F(2^{2^n})$ in $O(n)$ lines.

In this argument we won an exponential over the previous one. The price for this is that we implicitly used cuts and contractions to make the building blocks and to combine them. The proof of feasibility did not furnish a direct construction (of $2^{2^n}$). Concerning the contractions, to prove $F(x) \to F(x^2)$, one initially needs two copies of $F(x)$, and these are contracted into one.

We can win another exponential using quantifiers. That is, one can prove the feasibility of $2^{2^{2^n}}$ in $O(n)$ lines. The proof is constructed from the following building blocks. First we have that 2 is feasible, as above. Next we have that

$$(3) \qquad \forall x(F(x) \to F(x^2))$$

from the $F$ : *times* axiom. This is the same as before, except for the addition of the quantifier. The last building block is

$$(4) \qquad \forall x(F(x) \to F(x^k)) \to \forall x(F(x) \to F(x^{k^2}))$$

That is, we can use $\forall x(F(x) \to F(x^k))$ twice, the second time replacing $x$ with $x^k$, to get $\forall x(F(x) \to F(x^{k^2}))$. This is much better than in (2), since we are squaring the exponent instead of multiplying it by 2. By combining a series of these last building blocks, with $k = 2^{2^j}$, $j = 0, 1, \ldots, n - 1$, we can conclude that

$$(5) \qquad \forall x(F(x) \to F(x^2)) \to \forall x(F(x) \to F(x^{2^{2^n}})).$$

This can be combined with (3) to get a proof of $F(2^{2^{2^n}})$ in $O(n)$ lines.

This approach has some interesting features. As observed in (Carbone 2000a), the $F$ : *times* axiom is used only once, in the proof of (3). The proof of (4) is simply based on a substitution, and does not make use of any axiom for $F$. A contraction is also used in the proof of (4), in an interesting way; one employs $\forall x(F(x) \to F(x^k))$ twice to get $F(x) \to F(x^k)$ and $F(x^k) \to F((x^k)^k)$, and the two copies of $\forall x(F(x) \to F(x^k))$ can be contracted into each other. This is a standard point about quantifiers and contractions. They permit two occurrences of a formula to be contracted into one, even though the formulae have very different histories within the proof. More precisely, the quantifiers help to obtain formulae which are the same, and hence can be contracted, even if they come from formulae which are different, because of different terms inside. This kind of substitution did not occur in the previous propositional argument.

Notice that in this proof we did not have nesting of quantifiers. There are more elaborate proofs, due to Solovay, which use many nested quantifiers to get short proofs of very large numbers defined through towers of exponentials of arbitrary height. One gains an extra exponential with each nested quantifier. See (Carbone 2000a) for more information, and aspects of the dynamical structure of these proofs.

With these examples in mind, let us think about the type of description of a number provided by a proof of feasibility. In a proof of $F(2^n)$ in about $n$ lines using multiplications, we really make an explicit construction. We cannot expect to do better than win an exponential, because our most powerful operation is multiplication.

The other arguments are increasingly less explicit, with different types of combinations or substitutions. There is a kind of balance in this; as the proofs become shorter, their internal structure becomes more complicated, and there is increasing difficulty in the unwinding of implicit descriptions into explicit constructions. The internal structure of the proofs is more complicated both in terms of nested quantifiers, and in terms of nested cycles in the logical flow graphs of the proofs. See (Carbone 2000a; Carbone 2000c).

In giving short proofs of feasibility of large numbers like $2^{2^n}$ or $2^{2^{2^n}}$, we are using the special structure of these numbers, a kind of internal symmetry to them. This internal symmetry is reflected in the existence of short proofs, but there are no theorems about this. In general we should not be able to win so much compression using cuts, because arbitrary numbers will not have so much internal symmetry.

The mathematical idea of feasibility provides a way to embed arithmetic inside a space of formulae. Formal proofs then lead to new structure for natural numbers. This structure is quite different from the ones that are usually considered, and the cut rule plays an important role in this.

## 6. GROUPS

In recent years, much attention has been devoted to the study of the structure in finitely-generated groups which can be seen through the word metric. (See (Gromov 1993), for instance.) One fixes a generating set, and defines the distance from an element $g$ of the group $G$ to the identity $e$ to be the minimal length of the word that represents $g$. This can be extended to a left-invariant metric on all of $G$.

We can try to make other kinds of uniform measurements in the theory of groups using proofs and the idea of feasibility. Again let us introduce a unary predicate $F$, applied now to elements of our given group $G$. Let us also fix a finite subset $S$ of $G$ – we can think of it as a generating set, but actually the concept makes sense in any case – and require that $F$ have the following properties:

$$\begin{array}{ll} & F(e) \\ & F(\gamma) \quad \text{for each } \gamma \in S \\ F : equality & x = y \to (F(x) \to F(y)) \\ F : composition & F(x) \wedge F(y) \to F(xy) \\ F : inverse & F(x) \to F(x^{-1}) \end{array}$$

Here we write $xy$ for the group composition and $x^{-1}$ for the group inverse.

The length of the shortest proof of the feasibility of an element of $G$ can be taken to be some kind of measurement of its complexity. It is a well-defined function on $G$ because of the $F : equality$ rule. The length is always bounded by a constant multiple of the distance to $e$ in the word metric. We can make examples of proofs of feasibility which parallel the ones in the previous section. If $x \in G$ is feasible, then we can make a proof of $F(x^n)$ in $O(n)$ lines, by repeated use of the $F : composition$ rule. We can be more careful and get a proof of $F(x^{2^n})$ in $O(n)$ lines by making proofs of

$$(6) \qquad F(x^{2^j}) \to F(x^{2^{j+1}})$$

for $j = 0, 1, 2, \ldots, n-1$ as in (2) and combining them. This argument requires only propositional logical rules. If we use also quantifiers, then we can get a proof of $F(x^{2^{2^n}})$ in $O(n)$ lines as before. That is, the proof that we outlined before for (4) works just as well here.

The last method that we mentioned in Section 5, based on nesting of quantifiers, does not work directly in the theory of groups. To apply it to get a universal nonelementary distortion in groups (i.e, short proofs of $F(b) \to F(b^N)$ with $N$ a tower of exponentials like $2^{2^{2^{2^2}}}$), we would need to permit ourselves to quantify over integers as well as group elements. Indeed, for this argument we need to make substitutions into exponents, and this means substitutions with integers. The other arguments require only substitutions of group elements.

At any rate, we can make short proofs of feasibility using the first two methods. Given a finitely-generated group, these methods can be combined with the cancellation induced by the relations to yield even shorter proofs of feasibility. For example, let $G$ be the (Baumslag-Solitar) group with generators $x$ and $y$ and the single relation $y^2 = xyx^{-1}$. Thus $y^{2^m} = x^m yx^{-m}$. The feasibility of $x^m$ implies that of $x^m yx^{-m}$, and combining this with the earlier arguments one can get a proof of the feasibility of the group element $y^{2^{2^{2^n}}}$ in $O(n)$ lines. (See (Gromov 1993) for other examples of finitely-presented groups with distortion.)

Although in a sense we are simply transferring the earlier arguments for integers (from Section 5) to the theory of groups, there is an important difference between the two situations. In groups there are many ways to go to infinity. In a free group, for instance, every infinite (reduced) word describes a path to infinity in the associated tree. (A reduced word is one in which a generator $x$ never precedes or follows its inverse $x^{-1}$.) Our arguments about the integers lead to a lot of compression for proofs of feasibility along the direction of a cyclic subgroup $\{a^n\}$, at least for some $n$'s. In the word metric, all directions towards infinity in the free group are practically the same, but in the geometry of feasibility the cyclic subgroups are very special compared to generic directions. One can think of feasibility as providing a way to measure the amount of algebraic structure in a given direction.

This point can be seen in broader terms. The amount of compression that one can get for a notion of feasibility in some context can be seen as a measurement of the internal structure of the object in question. The examples in Section 5 reflect the internal symmetry in the case of arithmetic.

In the spirit of the recent theory of automatic groups (Epstein et al. 1992), one can be interested in representing a group through its set of words. We can enhance the notion of feasibility to be sensitive to the different ways that a group element is represented by words. Suppose now that our group $G$ is finitely presented, with a finite set $R$ of relations $w_i = e$ which express the triviality of the words $w_i$. We can introduce a new unary predicate $T$ on words so that $T(w)$ is intended as meaning that $w$ represents a trivial word. We impose the following axioms:

$$\begin{array}{ll}
 & T(e) \\
 & T(w_i) \quad \text{for each } w_i \in R \\
T : equality & w = u \to (T(w) \to T(u)) \\
T : composition & T(w) \wedge T(u) \to T(wu) \\
T : inverse & T(w) \to T(w^{-1}) \\
T : conjugation & T(w) \to T(vwv^{-1})
\end{array}$$

The idea of the last rule is that a trivial word conjugated by any word should again be trivial. It seems reasonable to make this rule without requiring that $v$ be feasible, but one might want to make different choices, depending on the situation.

For the purposes of making measurements in groups, one can combine the axioms for $F$ and $T$ and also add

$$F(w) \wedge T(u) \to F(wu) \quad \text{and} \quad F(w) \wedge T(u) \to F(uw).$$

Now $F$ is defined over words instead of group elements.

The notions presented in this section are not necessarily canonical or fixed. For instance, one might want to study chains of subgroups, each normal in the larger one, with different predicates for the different subgroups, each predicate axiomatized as above.

The bottom line is that proofs provide a nice way to try to look inside groups, to move around inside them and test their structure. This is an idea that has not been explored.

One feature of the idea of feasibility is its universality. It applies to all groups at once, like the word metric. This universality continues to exist under restrictions on the kind of proofs that we allow. This is an important point: one is free to choose a fragment of logic to suit one's purposes. Different fragments can lead to different metrics on groups.

## 7. RATIONAL NUMBERS

We can extend the idea of feasibility to rational numbers in a natural way. For our purposes, it will be convenient to consider $\infty$ as a rational number, with the conventions that $\infty \cdot \infty = \infty$, $a \cdot \infty = \infty \cdot a = \infty$ when $a \neq 0$, $0 \cdot \infty = \infty \cdot 0 = 0$, and $\frac{a}{\infty} = 0$ when $a$ is a finite rational number. We leave all other cases undefined. The need for $\infty$ is slightly a nuisance, but the point of it will be clear in a moment, and these technicalities are not serious.

We can introduce a feasibility predicate $F$ in much the same way as before. Now we want to measure "rational" complexity, and take the field structure into account. We use the same kind of axioms for $F$ as before, namely, that 0 and 1 are feasible, that equality, sums, and products preserve feasibility, and that additive and multiplicative inverses preserve feasibility. There are some small caveats needed to account for the cases when the operations are not defined.

We have seen how the feasibility of large integers can be established through short proofs, and we can do the same for rational numbers. This suggests natural open questions: how can one relate number-theoretic properties of a rational number to sizes of proofs of feasibility of it? As for groups, the restriction to different fragments of logic can lead to different properties, and one is free to choose logical systems to suit one's purposes.

Let us describe an amusing construction for feasibility of rational numbers. The basic point is that $2 \times 2$ matrices with (finite) rational entries act on rational numbers in a natural way. Let $\binom{a\,b}{c\,d}$ be such a matrix, and consider the transformation

$$(7) \qquad x \mapsto \frac{ax+b}{cx+d}$$

We assume that the determinant of our matrix is different from zero to avoid problems with the definition. This condition ensures that the numerator and the denominator above cannot both vanish at the same time, so that the quotient is always defined. It is for this reason that we allow $\infty$ as a rational number. If $x = \infty$, then we interpret the above quotient as being $\frac{a}{c}$. Not both of $a$ and $c$ can vanish, because of the assumption of nonzero determinant.

Let $A$ denote such a matrix $\binom{a\,b}{c\,d}$, and let $A$ also denote the projective linear transformation defined in (7). The correspondence from matrices to projective transformations is a homomorphism, i.e., products of matrices correspond to compositions of projective transformations. This is well known, and can be seen as follows. In working with the rational numbers $\mathbf{Q}$ together with $\infty$, we are really working with the projective line over the rational numbers, which means the space of ordinary (rational) lines in $\mathbf{Q} \times \mathbf{Q}$ that pass through the origin. If $\alpha \in \mathbf{Q}$, then we associate to it the line that passes through $(\alpha, 1)$. This parameterizes all lines in $\mathbf{Q} \times \mathbf{Q}$ through the origin except for the one which passes through $(1, 0)$, which we associate to $\infty$. A matrix $A$ with rational entries and nonzero determinant acts linearly on $\mathbf{Q} \times \mathbf{Q}$ and induces a transformation on the space of lines in $\mathbf{Q} \times \mathbf{Q}$ that pass through the origin. If we parameterize lines by rational numbers $\alpha$ together with $\infty$, then this induced transformation is the same as (7).

Let $x$ be a rational number, and consider $A^n x$. We would like to have short proofs of feasibility of $A^n x$ for large values of $n$. This fits with the earlier discussion for groups, i.e., since the collection of $2 \times 2$ matrices with rational entries and nonzero determinant forms a group. We can encode feasibility for the group of matrices in terms of feasibility for rational numbers. Given a $2 \times 2$ matrix $A = \binom{a\,b}{c\,d}$ with finite rational entries, let us write $\phi(A)$ for the formula $F(a) \wedge F(b) \wedge F(c) \wedge F(d)$. This extension of feasibility is preserved by matrix multiplication, by an easy argument. (Similarly, one could look at feasibility of rational numbers in terms of feasibility for integers, via feasibility of the numerator and denominator in a quotient of integers.)

This permits us to make short proofs of $\phi(A^n)$ for large values of $n$ as in Section 6. For the argument mentioned in Section 5 based on nested quantifiers there are some subtleties. To make the argument using nested quantifiers, one needs to quantify over integers (which would arise in the exponents of the matrix). One can do this if one can *define* the integers inside the field. For the rationals there is a way to do this, due to Julia Robinson (Robinson 1949). This would not work in a field of finite characteristic.

Once we have short proofs of $\phi(A^n)$ for large $n$, we can get short proofs of the feasibility of $A^n x$ for large $n$, given the feasibility of $x$. One could also look at other combinations of matrices, instead of powers of a single matrix. This is similar to the situation in Section 6.

We chose this example in part because of the well-known role of projective transformations in analysis and number theory. Let us briefly review some aspects of complex analysis and its connection with rational numbers. In complex analysis one works with complex numbers, both as matrix entries and for the domain on which the projective transformations act. Instead of having them act on the whole complex plane, one often restricts oneself to actions on the upper half-plane

$$\{z \in \mathbf{C} : z = x + iy,\ \ x, y \in \mathbf{R}, y > 0\}.$$

A well known corollary of the uniformization theorem (Ahlors 1973; Alhfors and Sario 1960) in complex analysis implies that most Riemann surfaces can be realized as the quotient of the upper half-plane by a discrete group of projective linear transformations. “Most” means all Riemann surfaces except the sphere, the plane, the plane with one puncture, and tori.

These group actions on the upper half-plane induce group actions on the boundary, the real line, which should be completed by the addition of a point at infinity. The action on the boundary can be much more chaotic than in the interior, with the orbit of a point being dense instead of discrete.

Sometimes Riemann surfaces and the corresponding groups of projective transformations have additional arithmetical structures. It can be natural to look at the action of the groups on rational numbers.

Here is a special case which provides an important example. Let $SL(2, \mathbf{Z})$ denote the group of 2 × 2 matrices with integer entries and determinant 1. This is indeed a group. The main point is that the inverses of such matrices still have integer entries, because the determinant is equal to 1. In fact, this group is well known to be finitely generated with generators $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. (See (Lang 1987, 30).) This group acts on the upper half-plane by projective transformations, and the quotient that results is isomorphic as a Riemann surface to the twice-punctured plane $\mathbf{C}\backslash\{0, 1\}$.

For the purpose of making rational numbers which admit short proofs of their feasibility, we can extend the preceding construction. The correspondence between matrices and projective transformations is a homomorphism, so that compositions of projective transformations correspond to products of matrices, and we can work directly with square matrices of any rank. Suppose that $A$ is an $m \times m$ matrix with rational entries. As before we can define a formula $\phi(A)$ which expresses the feasibility of the entries of $A$, and is it easy to show that $A \cdot B$ has feasible entries if $A$ and $B$ do. We can make short proofs of the feasibility of matrices $A^n$ for large $n$, in the same manner as discussed previously. The entries of $A^n$ are rational numbers which have short proofs of feasibility. One can also consider other combinations of matrices, as in the discussion in Section 6.

One can look at some of these matters more broadly in terms of dynamical processes. In proofs there is often a kind of dynamics going on, with various substitutions taking place or being described implicitly. Conversely, one might start with some type of dynamical system, as in the example of projective transformations acting on the rationals, and go from this to proofs with substantial structure.

## 8. A STORY FROM TOPOLOGY

The examples so far illustrate how one might use the idea of feasibility to make measurements and descriptions of mathematical constructions through proofs. The examples all had a kind of discreteness to them, and we would like to consider now a more “continuous” setting. For this we shall continue to not worry too much about formalization (and more than usual). Let us note, however, that although continuous notions can be convenient or desirable in some ways, one can often work just as well, or nearly so, with finite versions (e.g., using polyhedra and piecewise-linear mappings).

The concept of feasibility has a certain affinity with continuity. One could say that it seeks or entails a kind of connectedness.

Our example from topology will take some time to explain, and so we describe some general points first. We shall begin by reviewing the concept of *Serre fibrations* from topology (Serre 1951; Bott and Tu 1982). This notion involves the construction of continuous families of mappings for which the idea of feasibility can be relevant. To bring out this point, we shall discuss in some detail a particular example of a *torus bundle*. In this special case, the required construction amounts to taking large powers of a matrix in $SL(2, \mathbf{Z})$. More generally, in situations with smoothness, one can make constructions by solving ordinary differential equations, which can be seen as a continuous relative of taking large powers of a matrix. (We shall return to this in Section 9.)

This example of a torus bundle captures geometrically a basic phenomenon in proofs. Sometimes complicated constructions can be coded in short proofs through repeated cycling and substitutions. A proof may describe a simple operation which is used repeatedly in the actual construction. In our topological example, we shall see that the simple motion of cycling around a circle many times induces a motion up in our torus bundle with exponential distortion.

Let us now proceed with the details. Let $E$ and $B$ be two topological spaces, and let $\pi : E \to B$ be a continuous mapping between them. We say that $\pi : E \to B$ is a *Serre fibration* if it enjoys the following property. Let $P$ be a finite polyhedron, and suppose that we have continuous mappings $f : P \to E$ and $g : P \times [0, 1] \to B$ such that $\pi \circ f = g(\cdot, 0)$. Then there should be a "lifting" $\widehat{g} : P \times [0, 1] \to E$ of $g$, meaning a continuous mapping with $\pi \circ \widehat{g} = g$, such that $\widehat{g}(\cdot, 0) = f$. In other words, $g$ controls what happens in the base space $B$ over the whole interval $[0, 1]$, $f$ gives a compatible set of initial values in $E$ (for the point 0 in $[0, 1]$), and $\widehat{g}$ is a lifting of $g$ to $E$ which agrees with the initial values specified by $f$. This is similar to the lifting of paths in covering surfaces (Ahlfors and Sario 1960; Massey 1991), but now we are working with continuous families of paths parameterized by the polyhedron $P$. (Lifting a single path would correspond to taking $P$ to be a single point.)

To understand what this means, consider the simple case where $E = B \times F$ for some topological space $F$ (which gives the "fibers"). In this event, the fibration property is automatic, and one can write down a choice of $\widehat{g}$ directly. Namely, one can take $\widehat{g}(p, x) = (g(p, x), \pi_1(f(p)))$, where $\pi_1 : E \to F$ is the obvious projection onto $F$. Fibrations of this form are called *trivial*. A fundamental class of nontrivial fibrations consists of ones in which $E$ looks like a product above small open sets in the base $B$ (i.e., the fibration is locally trivial on $B$), but for which there is nontrivial twisting globally on $B$. In these cases, one can often verify the existence of the necessary liftings by exhausting $P \times [0, 1]$ through local liftings. Compactness assumptions are frequently used in this regard, to ensure that the exhaustion works.

Let us think of the lifting $\widehat{g} : P \times [0, 1] \to E$ of $g$ above as being like an explicit proof of feasibility. We start with an initial configuration which is given by $f : P \to E$, and $\widehat{g}$ provides a way to get from $f$ to a final configuration given by $\widehat{g}(\cdot, 1)$ in a continuous manner. In a discrete setting, one would think of a sequence of small or simple steps. Normally, there can be a definite amount of distortion at each step, and then expo-

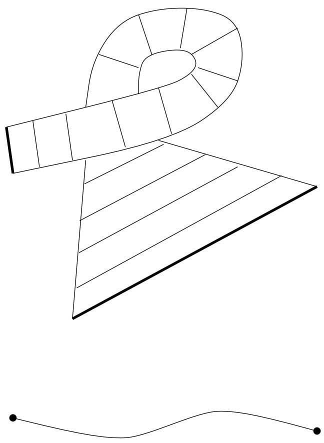

*Figure 1.*

nential distortion over the whole time interval [0, 1]. See Figure 1 for an illustration related to this.

To make the notions more clear, let us consider a concrete example of a fibration with nontrivial twisting. Let $\mathbf{S}^1$ denote the unit circle. It will be more convenient to think of it as the quotient space $\mathbf{R}/\mathbf{Z}$. Let $T$ denote the torus $\mathbf{S}^1 \times \mathbf{S}^1$, which we can think of as $\mathbf{R}^2/\mathbf{Z}^2$.

We want to look at *torus bundles* over a circle. We shall use the following recipe. Suppose that $A : T \to T$ is a homeomorphism. Take $[0, 1] \times T$ and glue the two ends $\{0\} \times T$ and $\{1\} \times T$ together using $A$. This means that we take $[0, 1] \times T$ and we identify $(0, u)$ with $(1, A(u))$ for all $u \in T$. This defines a space which we call $E$. There is a natural mapping $\pi : E \to \mathbf{S}^1$ which corresponds to the projection of $[0, 1] \times T$ onto $[0, 1]$, where we identify $\mathbf{S}^1$ with the space obtained by taking $[0, 1]$ and identifying the endpoints 0 and 1.

We can describe this space in another way as follows. We start with $\widetilde{E} = \mathbf{R} \times T$. We define a mapping $\phi : \widetilde{E} \to \widetilde{E}$ by $\phi(x, u) = (x + 1, A(u))$. This mapping generates an infinite cyclic group of homeomorphisms on $\widetilde{E}$, and $E$ is just the quotient of $\widetilde{E}$ by this group, in the same way that $\mathbf{S}^1$ is the quotient of $\mathbf{R}$ by the infinite cyclic group of homeomorphisms generated by $x \mapsto x + 1$.

If instead of $\phi$ we used the mapping $(x, u) \mapsto (x + 1, u)$ we would simply get $\mathbf{S}^1 \times T$ for the quotient. By choosing a suitable mapping $A : T \to T$ we can get a bundle in which there is some nontrivial twisting as we go around the base.

Let us consider now a specific example of such a mapping $A$. Start with the matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, which lies in $SL(2, \mathbf{Z})$ since it has determinant 1. This defines a linear mapping on $\mathbf{R}^2$. Because the matrix has integer entries, the corresponding linear mapping sends the standard integer lattice $\mathbf{Z}^2$ inside $\mathbf{R}^2$ to itself. Thus we can get a well-defined mapping on the quotient $\mathbf{R}^2/\mathbf{Z}^2 = T$, and we take this to be $A$. This defines a homeomorphism on $T$, because the inverse of $\binom{2\,1}{1\,1}$ is also a matrix with integer entries (since the determinant is 1), and hence it descends to a mapping on $T$ as well.

Thus we get a homeomorphism $A : T \to T$. It may seem harmless, but in fact it is quite nontrivial. It is not homotopic to the identity, for instance. For if it were, its lifting to the universal covering of $T$ would differ from the identity mapping by only a bounded amount, and this is not true. Indeed, $\mathbf{R}^2$ is the universal covering of $T$, and the lifting of $A$ to it is the linear transformation with which we started.

The first homology and homotopy groups of $T$ are isomorphic to each other and to $\mathbf{Z}^2$. By general facts, the homeomorphism $A$ induces an automorphism on this group, which in this case is given by the action of the matrix $\binom{2\,1}{1\,1}$ with which we began. This provides a topological way to measure the difference between $A$ and the identity mapping, even up to homotopy. That is, the identity mapping on $T$ induces, in the same manner, the identity mapping on $\mathbf{Z}^2$, and these induced mappings are preserved by homotopies of the original mappings on $T$.

To understand better the nontrivial effect of $A$, it is helpful to compute the eigenvalues of the matrix $\binom{2\,1}{1\,1}$. These are the roots of the polynomial

$$\det\left(\begin{pmatrix}2&1\\1&1\end{pmatrix} - \lambda I\right) = (2-\lambda)(1-\lambda) - 1 = \left(\lambda - \frac{3}{2}\right)^2 - \frac{5}{4},$$

namely $\lambda = \frac{3\pm\sqrt{5}}{2}$. Note that the product of these numbers is 1, as it should be, and that one is larger than 1 and the other is smaller than 1. In fact the larger eigenvalue is between 2 and 3.

Because the matrix is symmetric, we can find an orthogonal basis with respect to which it is diagonal with these two eigenvalues. When we take large powers of the matrix we get exponential compression in one direction and exponential expansion in the other. In topological terms, this exponential expansion for the matrix implies that there are loops in the torus $T$ whose image in $T$ under $A^n$ wraps around an exponentially larger number of times. This wrapping is depicted by the diagonal lines in Figure 2, where the torus is obtained from the square by identifying the opposite sides. The diagonal lines represent a single curve in the torus, which goes across the square several times.

Thus $A$ is quite nontrivial and this leads to nontrivial twisting of the fibration $E$. We want to see how this twisting is reflected in liftings (as for Serre fibrations). Suppose that $f : P \to E$ is some continuous mapping, where $P$ is a finite polyhedron. For simplicity, assume that the image of

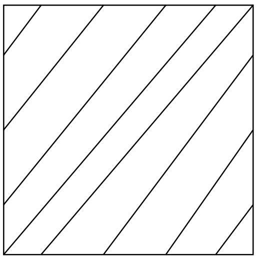

*Figure 2.*

$f$ lies in a single fiber of $E$, so that there is a point $b \in \mathbf{S}^1$ such that the image of $f$ lies in $\pi^{-1}(b)$. Let us assume also that our mapping $g$ is of a particularly simple form, that it is constant on $P$. Thus $g$ is in essence a mapping from [0, 1] into $\mathbf{S}^1$, which we denote by $\gamma$, since it is technically a separate object (a mapping on [0, 1] instead of $P \times [0, 1]$). Note that $\gamma(0) = b$, because of the compatibility between $g$ and $f$. In this situation, our lifting problem becomes that of finding a mapping $f_1 : P \times [0, 1] \to E$ which is an extension of $f$, in the sense that $f_1(p, 0) = f(p)$ for all $p \in P$, and whose projection to the base is essentially $\gamma$, in the sense that $\pi(f_1(p, x)) = \gamma(x)$ for all $x \in [0, 1]$ and all $p \in P$.

In other words, $f$ maps $P$ into a single fiber of $E$, and $f_1(\cdot, x)$ should map $P$ into the fiber in $E$ over $\gamma(x)$ for each $x$. As $x$ ranges through [0, 1], these fibers can move, and one returns to the same fiber when $\gamma$ contains loops.

Let us explain how we can obtain such a lifting $f_1(p, x)$. If $E$ were just the product $\mathbf{S}^1 \times T$, then we could pull $f$ along the parameter interval rigidly, as we discussed earlier in the section. Because of the twisting of our bundle, we have to do something more complicated, and it is convenient to go back to $\widetilde{E} = \mathbf{R} \times T$. Let $\widetilde{\gamma} : [0, 1] \to \mathbf{R}$ be a continuous mapping which projects to $\gamma$ under the canonical mapping from $\mathbf{R}$ to $\mathbf{R}/\mathbf{Z} = \mathbf{S}^1$. This lifting $\widetilde{\gamma}$ of $\gamma$ is determined uniquely by its initial point $\widetilde{b} = \widetilde{\gamma}(0)$, which is a lifting of $b$ (and otherwise arbitrary). Our quotient mapping from $\widetilde{E}$ onto $E$ is a homeomorphism on each of the fibers, and hence there is a mapping $\widetilde{f} : P \to \widetilde{E}$ which takes values in $\{\widetilde{b}\} \times T$ and which projects down to $f$ when we project $\widetilde{E}$ onto $E$ using our quotient mapping.

In short, we can lift $\gamma$ and $f$ upstairs to $\widetilde{E}$, and in a compatible way. Since $\widetilde{E} = \mathbf{R} \times T$, we can define a mapping $\widetilde{f_1} : P \times [0, 1] \to \widetilde{E}$ in the obvious way, by setting

$$\widetilde{f_1}(p, x) = (\widetilde{\gamma}(x), \widetilde{\pi}_1(\widetilde{f}(p))),$$

where $\widetilde{\pi}_1 : \widetilde{E} \to T$ is the standard projection. Thus, up in $\widetilde{E}$, we are doing something quite trivial, we are simply sliding $\widetilde{f}$ along $\mathbf{R}$ rigidly.

Define $f_1 : P \times [0, 1] \to E$ to be the composition of $\widetilde{f_1} : P \times [0, 1] \to \widetilde{E}$ with the quotient mapping from $\widetilde{E}$ onto $E$. It is easy to see that this choice of $f_1$ has the desired properties, namely, that it is a continuous mapping which agrees with $f$ at the beginning, and follows $\gamma$ in the base for the whole time interval [0, 1].

Now let us look at how $f_1(p, x)$ is distorted as $x$ goes from 0 to 1. Imagine that $\gamma$ moves at constant speed in $\mathbf{S}^1$ and reasonably quickly, so that it wraps around $\mathbf{S}^1$ numerous times. To be specific, assume that $\gamma$ wraps around $\mathbf{S}^1$ exactly $n$ times, moving in the positive orientation and ending back at its initial point $b$. The lifting $\widetilde{\gamma}$ of $\gamma$ to $\mathbf{R}$ moves along at constant speed as well, starting at $\widetilde{b}$ and ending at $\widetilde{b} + n$.

What does $f_1(\cdot, 1)$ look like when we go around the circle $n$ times? It looks like $f_1(\cdot, 0) = f$ acted on by $A^n$! That is, $f_1(\cdot, 1)$ and $f$ both map $P$ into the fiber $\pi^{-1}(b)$ in $E$, which is a copy of $T$. If we move the mappings back into $T$ so that we can look at them, then the transition from time 0 to time 1 is given by $A^n$. This is because each tour around $\mathbf{S}^1$ corresponds to an application of $A$ on $T$. This follows from the definitions.

From our earlier analysis of $A$, we conclude that our mapping $f_1$ may undergo exponential stretching as we traverse the parameter interval [0, 1]. This is unavoidable, and not simply an artifact of the particular construction of $f_1$. For instance, suppose that our initial mapping $f$ represents a loop in $T$ (i.e., $P$ is a polygonal loop). The ending mapping $f_1(\cdot, 1)$ represents another loop in $T$. No matter how we choose $f_1$, the homotopy class of our final loop in $T$ has to be the same as the one obtained from the construction above. This is a basic property of Serre fibrations. Therefore, the amount of winding in $T$ that the final loop makes in terms of topology is simply determined by $A^n$, and can be exponentially large, as we have seen.

This finishes our concrete construction. Let us think about what it means. Consider notions of feasibility for mappings into the space $E$. We might call such a mapping feasible if it is quite simple (e.g., if it does not wrap around too much), or if it can be obtained from a feasible mapping by a small perturbation. Thus feasibility is like being homotopic to something simple. One can try to find complicated objects which are feasible with a short proof, and the preceding discussion suggests examples of this.

## 9. FEASIBILITY AND CONTINUOUS PARAMETERS

In ordinary mathematics one often works with continuous constructions, such as taking $B^t$ where $t$ is a real number, rather than an integer. One can do this with matrices, for instance, and something similar takes place in solving an ordinary differential equation. What about ideas of feasibility in situations like these?

As a basic example, consider the differential equation $y' = y$, which is solved by the exponential function $e^x$. One can view this differential equation as corresponding to a continuous version of recursion, and the existence of its solution as a consequence of a "continuous" version of induction. (One can think of this as being analogous to proving by induction in arithmetic that exponential functions are defined everywhere.) For the notion of feasibility in arithmetic, it was important not to allow induction over $F$-formulae, to avoid collapsing into triviality. A theory which provides the existence of exponential functions with continuous parameters, or solutions of differential equations, might arguably be too strong in a similar manner.

In ordinary mathematics, one might define the exponential function by summing an infinite series, or as the inverse of the natural logarithm which is defined using an integral, and one can find solutions of ordinary differential equations through conversions to integral equations and analysis there. As a more step-by-step approach, one can use the formula

$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

for the exponential, and obtain solutions of ordinary differential equations as limits of solutions of discrete difference equations. These different approaches can have advantages and disadvantages, depending on the context. It might be useful to work more "globally", with series or integrals, while step-by-step methods have other features.

What about the topological situation in the previous section? In many settings one can use solutions of ordinary differential equations to obtain liftings with properties as in Section 8. On the other hand, one does not need to know that the family of mappings being sought satisfies any particular differential equation (even if that might be a useful method). There is some extra flexibility, where certain kinds of perturbations do not cause trouble. Once one has an approximation which is sufficiently fine, and which can be discrete, it is easy to fill in or adjust the rest to have a continuous solution with the requisite properties.

Roughly speaking, one might say that for the topological situation in Section 8, $(1+\frac{\alpha}{n})^n$ can be practically as good as $e^\alpha$ when $n$ is large enough.

Of course, one can consider notions of feasibility for approximating classical constructions in analysis (such as summing a series for the exponential function, or approximation schemes for finding solutions of ordinary differential equations). For these types of constructions, one typically deals with efficiency of approximation rather than exact values, as in the earlier and more algebraic examples. In topology a good approximation is often good enough already.

## 10. SOME REMARKS ABOUT COMPRESSION AND CUTS

We have tried to explore some examples of natural notions of “feasibility”, where there can be short proofs.

What does “short” really mean? In the examples there was some clear sense that the proofs were short compared to what one might expect, given the complexity of the particular object in question. But can we define “short” more abstractly, more invariantly, more objectively?

An answer to this is given in terms of the cut rule in sequent calculus. The reader who is unfamiliar with this may wish to consult (Carbone and Semmes 1997) for an introduction to the cut rule and the combinatorics and complexity of cut elimination. (See (Girard 1987b; Takeuti 1975) for more extensive treatments.) The short proofs mentioned above all use cuts, and a natural way to measure the “shortness” is to ask how large a proof would have to be without the cut rule. We would then consider the *relative* sizes between proofs with and without cuts rather than absolute sizes of proofs.

More precisely, and as mentioned in Section 5, it is the interaction between cuts and contractions which gives rise to the shortness of the proofs. The elimination of cuts leads to simplification of the manner in which contractions can be used.

In many contexts in logic one can show that it is possible to eliminate cuts from a proof, but at potentially great cost of expansion. See (Gentzen 1934; Gentzen 1969; Girard 1987b; Takeuti 1975). There are examples known where the smallest proof without cuts is much larger than the smallest proof with cuts (Tseitin 1968; Orevkov 1982; Orevkov 1993; Statman 1974; Statman 1978; Statman 1979; Haken 1985; Buss 1987). The present discussion suggests that we view this phenomenon as a reflection of some kind of internal symmetry or structure. With notions of feasibility, formal proofs deal with mathematical objects and the building of them in a precise way, and one can consider the relationship between proofs and the internal symmetry or structure of the mathematical objects which are involved. (This is part of the reason for looking at feasibility, as we do here.) In particular, contractions play an important role, as before. See also (Carbone and Semmes 1999; Carbone and Semmes 2000) in connection with these topics. These issues are related as well to the “P = NP?” and “NP = co-NP?” problems (Cook and Reckhow 1979; Garey and Johnson 1979; Johnson 1990).

One might compare these matters with common mathematical experience, and the proofs that are made. The situations with which mathematicians deal typically involve a lot of special structure or symmetry (in some form), and this is not really accidental.

## ACKNOWLEDGMENTS

A. Carbone is supported by the Lise-Meitner Stipendium # M00187-MAT (Austrian FWF). Both authors are grateful to the Institut des Hautes Études Scientifiques (Bures-sur-Yvette, France) for its hospitality. S. Semmes is supported by the U.S. National Science Foundation.

## REFERENCES

Ahlfors, L.: 1973, *Conformal Invariants: Topics in Geometric Function Theory*, McGraw-Hill, New York.

Ahlfors, L. and L. Sario: 1960, *Riemann Surfaces*, Princeton University Press, Princeton.

Bott, R. and L. Tu: 1982, *Differential Forms in Algebraic Topology*, Springer-Verlag, Berlin.

Buss, S.: 1987, 'Polynomial Size Proofs of the Propositional Pigeonhole Principle', *Journal of Symbolic Logic* **52**, 916–27.

Buss, S.: 1991, 'The Undecidability of $k$-Provability', *Annals of Pure and Applied Logic* **53**, 75–102.

Carbone, A.: 1997, 'Interpolants, Cut Elimination and Flow Graphs for the Propositional Calculus', *Annals of Pure and Applied Logic* **83**, 249–99.

Carbone, A.: 2000a, 'Cycling in Proofs and Feasibility', *Transactions of the American Mathematical Society* **352**, 2049–2075.

Carbone, A.: 2000b, 'Some Combinatorics behind Proofs', (submitted).

Carbone, A.: 2000c, 'Asymptotic Cyclic Expansion and Bridge Groups of Formal Proofs', (submitted).

Carbone, A. and S. Semmes: 1997, 'Making Proofs Without Modus Ponens: An Introduction to the Combinatorics and Complexity of Cut Elimination', *Bulletin of the American Mathematical Society* **34**, 131–59.

Carbone, A. and S. Semmes: 1999, 'Propositional Proofs via Combinatorial Geometry and the Search for Symmetry', in *Collegium Logicum*, Annals of the Kurt-Gödel-Society, Volume 3, pp. 85–98, Institute for Computer Science AS CR, Prague.

Carbone, A. and S. Semmes: 2000, *A Graphic Apology for Symmetry and Implicitness – Combinatorial Complexity of Proofs, Languages, and Geometric Constructions*, forthcoming, Oxford University Press, Oxford.

Chaitin, G.: 1987, *Information, Randomness & Incompleteness – Papers on Algorithmic Information Theory*, World Scientific, Singapore.

Chaitin, G.: 1992, *Information-Theoretic Incompleteness*, World Scientific, Singapore.

Connes, A.: 1994, *Noncommutative Geometry*, Academic Press, New York.

Cook, S. and R. Reckhow: 1979, 'The Relative Efficiency of Propositional Proof Systems', *Journal of Symbolic Logic* **44**, 36–50.

Craig, W.: 1957, 'Three Uses of the Herbrand-Gentzen Theorem in Relating Model Theory and Proof Theory', *Journal of Symbolic Logic* **22**, 269–85.

Epstein, D., J. Cannon, D. Holt, M. Paterson, and W. Thurston: 1992, *Word Processing in Groups*, Jones and Bartlett, Boston.

Garey, M. and D. Johnson: 1979, *Computers and Intractability : A Guide to the Theory of NP-Completeness*, W. H. Freeman, New York.

Gentzen, G.: 1934, 'Untersuchungen über das Logische Schließen I-II', *Mathematische Zeitschrift* **39**, 176–210, 405–31.

Gentzen, G.: 1969, in M. Szabo (ed.), *The Collected Papers of Gerhard Gentzen*, North-Holland, Amsterdam.

Girard, J.-Y.: 1987a, 'Linear Logic', *Theoretical Computer Science* **50**, 1–102.

Girard, J.-Y.: 1987b, *Proof Theory and Logical Complexity*, Bibliopolis, Napoli.

Girard, J.-Y.: 1989a, 'Towards a Geometry of Interaction', in J. Gray and A. Scedrov (eds), *Categories in Computer Science and Logic*, American Mathematical Society, Providence, pp. 69–108.

Girard, J.-Y.: 1989b, 'Geometry of Interaction I: Interpretation of System *F*', in R. Ferro, C. Bonotto, S. Valentini, and A. Zanardo (eds), *Logic Colloquium '88*, North Holland, Amsterdam, pp. 221–60.

Girard, J.-Y.: 1990, 'Geometry of Interaction II: Deadlock-Free Algorithms', in P. Martin-Löf and G. Mints (eds), *COLOG-88*, Springer-Verlag, Berlin, pp. 76–93.

Girard, J.-Y.: 1995, 'Geometry of Interaction III: Accommodating the Additives', in J.-Y. Girard, Y. Lafont, and L. Regnier (eds), *Advances in Linear Logic*, Cambridge University Press, Cambridge, pp. 329–89.

Gromov, M.: 1993, 'Asymptotic Invariants of Infinite Groups', in G. Niblo and M. Roller (eds), *Geometric Group Theory*, Volume 2, Cambridge University Press, Cambridge, pp. 1–295.

Haken, A.: 1985, 'The Intractability of Resolution', *Theoretical Computer Science* **39**, 297–308.

Halmos, P.: 1974, *Lectures on Boolean Algebras*, Springer-Verlag, Berlin.

Johnson, D.: 1990, 'A Catalog of Complexity Classes', in J. van Leeuwen (ed.), *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, Elsevier and MIT Press, Amsterdam and Cambridge, MA, pp. 67–161.

Kirwan, F.: 1992, *Complex Algebraic Curves*, Cambridge University Press, Cambridge.

Kolmogorov, A.: 1968, 'Logical Basis for Information Theory and Probability Theory', *IEEE Transactions on Information Theory* **14**(5), 662–64.

Lang, S.: 1987, *Elliptic Functions*, Springer-Verlag, Berlin.

Li, M. and P. Vitányi: 1990, 'Kolmogorov Complexity and its Applications', in J. van Leeuwen (ed.), *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, Elsevier and MIT Press, Amsterdam and Cambridge, MA, pp. 187–254.

Massey, W.: 1991, *A Basic Course in Algebraic Topology*, Springer-Verlag, Berlin.

Orevkov, V.: 1982, 'Lower Bounds for Increasing Complexity of Derivations after Cut Elimination', *Journal of Soviet Mathematics* **20**, 2337–350.

Orevkov, V.: 1993, in A. Bochman (trans.), D. Louvish (ed.), *Complexity of Proofs and Their Transformations in Axiomatic Theories*, American Mathematical Society, Providence.

Parikh, R.: 1971, 'Existence and Feasibility in Arithmetic', *Journal of Symbolic Logic* **36**, 494–508.

Robinson, J.: 1949, 'Definability and Decision Problems in Arithmetic', *Journal of Symbolic Logic* **14**, 98–114.

Rudin, W.: 1991, *Functional Analysis*, McGraw-Hill, New York.

Serre, J.-P.: 1951, 'Homologie Singulière des Espaces Fibrés', *Annals of Mathematics* **54**, 425–505.

Simmons, G.: 1963, *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York.

Statman, R.: 1974, *Structural Complexity of Proofs*, doctoral dissertation, Stanford University.

Statman, R.: 1978, 'Bounds for Proof-Search and Speed-Up in the Predicate Calculus', *Annals of Mathematical Logic* **15**, 225–87.

Statman, R.: 1979, 'Lower bounds on Herbrand's Theorem', *Proceedings of the American Mathematical Society* **75**, 104–7.

Stein, E.: 1970, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton.

Takeuti, G.: 1975, *Proof Theory*, North-Holland, Amsterdam.

Tseitin, G.: 1968, 'Complexity of a Derivation in the Propositional Calculus', *Zap. Nauchn. Sem. Leningrad Otd. Mat. Inst. Akad. Nauk SSSR* **8**, 234–59.

Urquhart, A.: 1987, 'Hard Examples for Resolution', *Journal of the Association for Computing Machinery* **34**, 209–19.

van Leeuwen, J. (ed.): 1990, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, Elsevier and MIT Press, Amsterdam and Cambridge, MA.

A. Carbone
Department of Mathematics and Computer Science
University of Paris 12
61 Avenue du Général de Gaulle
94010 Créteil cedex
France

S. Semmes
Department of Mathematics
Rice University
Houston, Texas, 77251
U.S.A.